Regression Tree-Based Methodology for Customizing Building Energy Benchmarks to Individual Commercial Buildings

by

Apoorva Prakash Kaskhedikar

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved July 2013 by the
Graduate Supervisory Committee:

T. Agami Reddy, Chair
Harvey Bryan
George Runger

ARIZONA STATE UNIVERSITY

August 2013

ABSTRACT

According to the U.S. Energy Information Administration, commercial buildings represent about 40% of the United State's energy consumption of which office buildings consume a major portion. Gauging the extent to which an individual building consumes energy in excess of its peers is the first step in initiating energy efficiency improvement. Energy Benchmarking offers initial building energy performance assessment without rigorous evaluation. Energy benchmarking tools based on the Commercial Buildings Energy Consumption Survey (CBECS) database are investigated in this thesis.

This study proposes a new benchmarking methodology based on decision trees, where a relationship between the energy use intensities (EUI) and building parameters (continuous and categorical) is developed for different building types. This methodology was applied to medium office and school building types contained in the CBECS database. The Random Forest technique was used to find the most influential parameters that impact building energy use intensities. Subsequently, correlations which were significant were identified between EUIs and CBECS variables. Other than floor area, some of the important variables were number of workers, location, number of PCs and main cooling equipment. The coefficient of variation was used to evaluate the effectiveness of the new model.

The customization technique proposed in this thesis was compared with another benchmarking model that is widely used by building owners and designers namely, the

ENERGY STAR's Portfolio Manager. This tool relies on the standard Linear Regression methods which is only able to handle continuous variables. The model proposed uses data mining technique and was found to perform slightly better than the Portfolio Manager. The broader impacts of the new benchmarking methodology proposed is that it allows for identifying important categorical variables, and then incorporating them in a local, as against a global, model framework for EUI pertinent to the building type. The ability to identify and rank the important variables is of great importance in practical implementation of the benchmarking tools which rely on query-based building and HVAC variable filters specified by the user.

To Aai and Baba

ACKNOWLEDGEMENTS

I would like to express my gratitude to Professor T. Agami Reddy, my thesis Chair, for his constant encouragement and guidance throughout the thesis and the master's program at Arizona State University. I would take this opportunity to thank Prof. Harvey Bryan and Prof. George Runger for their invaluable suggestions throughout this research.

I am thankful to my colleague and friend Mr.Ranojoy Dutta for his valuable insights during the course of this project. I would also like to thank all my friends for their patience, encouragement, and constant support. Special thanks to my best friends Priya, Preethi and Vishal for standing by me always. I am immensely grateful to my parents, sister and family for being a source of strength at all times, in all my pursuits.

TABLE OF CONTENTS

Page

LIST OF TABLES viii

LIST OF FIGURES xi

CHAPTER

1. INTRODUCTION 1

1.1 Background 1

1.2 Problem Statement 5

1.3 Objective 7

1.4 Scope 7

2. LITERATURE REVIEW 8

2.1 Base lining 8

2.2 Benchmarking 8

2.3 Benchmarking tools 9

2.3.1 Energy Star Portfolio Manager 9

2.3.2 ASHRAE Building Energy Quotient (BEQ) 18

2.3.2 Energy IQ 22

2.3.3 ORNL Spreadsheet 25

2.3.4 FEDS (Retro-commissioning tool) 28

2.4 Model used for energy benchmarking 33

CHAPTER Page

2.4.1 Linear Regression Study 33

2.4.2 Simulation Study 35

3. METHODOLOGY 38

3.1 Selecting bldg type and variables 39

3.2 Normalizing the variables 40

3.3 Removing Outliers 40

3.4 Calculating the variability in EUI 42

3.5 Linear Regression 42

3.6 Data mining approach 45

3.6.1 Decision tree 46

3.6.2 Decision tree Algorithm 49

3.6.3 CART Performance Metrics 50

3.6.4 Random Forest 52

3.6.5 Advantages and Disadvantages of Decision trees 53

3.7 Reducing the categories in the variables/ Discretization 55

4. APPLICATION TO CBECS DATA 56

4.1 Commercial Buildings Energy Consumption Survey (CBECS) 56

4.1.1 Office 59

CHAPTER Page

4.1.2 School....66
4.2 Discretization methodology....67
4.3 Random Forest for office buildings....74
4.4 Combining Office and School data....75
4.5 Random Forest for combined building data....77
4.6 Identifying important Variables....79
4.7 Comparative analysis with prior work....83
4.8 Model Accuracy....88
4.8.1. For Office Buildings:....88
4.8.2. For Office and School Buildings (Combined Data)....88
5. SUMMARY AND FUTURE WORK....92
5.1 Summary....92
5.2 Limitations....93
5.3 Future Research....94
REFERENCES....95
APPENDIX A....100
APPENDIX B....113
APPENDIX C....118

LIST OF TABLES

Table Page

1. Descriptive Statistics for variables in the final OLS model (US EPA, 2007) 12

2. Computing Building Centered Variables (US EPA, 2007) 15

3.Computing Predicted Source EUI (US EPA, 2007)..... 16

4. List of energy efficiency ratio cut-off points for each rating, from 1 to 100 (US EPA, 2007)..... 17

5. Difference between Operational and Asset rating (ASHRAE Building Energy Labeling Program Implementation Committee, 2009) 20

6. Difference between Statistical and Technical rating (ASHRAE Building Energy Labeling Program Implementation Committee, 2009)..... 21

7. Rating and Energy Use indicators (ORNL,1996) 28

8. Altered building parameters in sensitivity analysis (Huang et.al, 1993)..... 36

9.Description of Variables 60

10. EUI for entire dataset for Office buildings 61

11. Final list of selected variables 62

12. Outliers detected for Office buildings 63

13. The Inter-quartile range for Office buildings..... 64

14. Descriptive Statistics of the range (Office buildings) 64

15. Descriptive Statistics after removal of outliers for Office buildings..... 65

16. Descriptive Statistics for School buildings without outlier removal..... 66

Table Page

17. Descriptive statistics for School buildings after outlier removal 67

18. Discretization for Office bldg. variables..... 68

19. Office buildings EUI range discretized in 3 and 5 categories..... 69

20. Threshold of each class of the variable for office buildings 71

21. Collapsing the classes for Office buildings 71

22. Description of each class for Office buildings 72

23. Collapsing the categories for Office buildings..... 72

24. Definition of classes for Working operating hours (in hours) for Office and School buildings 73

25. Collapsing the classes for Office and school buildings 73

26. Confusion Matrix for 3 categories..... 74

27. Confusion Matrix for 5 categories..... 75

28. Description statistics for combined data 76

29. Description of 5 categories for combined data 76

30. Description of 4 categories for combined data 77

31. Confusion matrix for the other method of classification into 5 categories..... 78

32. Confusion matrix for 4 categories 78

33. Variable Importance 79

34. Ranking of variables 81

35. Variable importance for regression ensemble 82

36. Variable importance for classification ensemble 83

Table Page

37. Normalizing the continuous variables (ENERGY STAR, Portfolio Manager) 84

38. Model Summary for Office Building data 84

39. Coefficients for the model 85

40. Model Summary for combined data 86

41. Coefficients for the model for combined data..... 86

42.Model summary for medium Office buildings (based on Sharp's model) 87

43. Coefficients for medium Office buildings (based on Sharp's model)..... 87

44. Comparison of Models 88

45. Comparison of models 89

LIST OF FIGURES

Figure Page

1: US energy consumption (US Department of Energy (DOE), 2008) 1

2: U.S. Building Impacts ("Green building alliance", 2013) 3

3. Cycle of Improvement (EEBHUB, 2013) 4

4.&5. Statistical Rating Scale & Technical Rating Scale 22

6. ASHRAE Building Energy Quotient Label 22

7. Action-Oriented Benchmarking (LBNL,2008) 23

8. Energy IQ Result view 25

9. ORNL Spreadsheet view 26

10. Baseline Methodology 39

11. Example of Outliers ("Mark Young Training Systems", n.d.) 41

12. Box-plot and probability density function of a normal distribution ("Inter-quartile range", 2013) 42

13. Simple linear regression (Wikipedia) 43

14. Decision tree format 47

15. Recursive Partitioning (Kong et al., 2012) 48

16. Black Box Model (Breiman, 2003) 52

17. Region and Census division Map (CBECS, 2003) 56

18. Climate Zones (CBECS, 2003) 57

19. Electricity Use vs. Reported Floor area (Sharp, 1996) 58

20. Office-Building distribution by square footage 59

Figure Page

21. Graph to show EUI distribution for office buildings 61

22. Outlier Detection based on EUI 63

23. EUI distribution for the entire data (Office buildings) 64

24. Graph for EUI distribution after the removal of outliers for Office buildings 65

25. EUI distribution (with the presence of outliers) for School buildings 66

26. Graph-EUI distribution for School buildings 67

27. Distribution for three categories for Office buildings 70

28. Distribution for five categories for Office buildings 70

29. Distribution of Glass % 71

30. Distribution of classes for bldg shape 72

31. Combining similar classes for combined dataset 73

32. Distribution of EUI with 5 categories for combined data 76

33. Distribution of EUI with 4 categories for combined data 77

34. Graph depicting the important variables of Office buildings (Regression version) ... 80

35. Graph depicting the important variables of Office buildings (Classification version) 81

36.Important variables in combined dataset (Regression version) 82

37.Important variables for combined dataset (Classification version) 83

38. Single regression tree for office buildings using 5 important variables 90

39. Single regression tree using 6 most important variables for office and School buildings combined 90

CHAPTER 1

INTRODUCTION

1.1 Background

Energy Information Administration (US EIA, 2012) in the Annual Energy Outlook 2012 states that the overall energy consumption in the US would grow at an average annual rate of 0.3% from 2010 to 2035. The projected energy demand for transportation is estimated to grow at an annual rate of 0.1% from 2010 through 2035, and electricity demand by 0.7% per year, primarily as a result of rising energy consumption in the buildings sector.

When classified into residential, commercial, transportation and industrial sectors, the largest increase, 7.2 quadrillion Btu from 2009 to 2035, is attributed to the industrial sector, which was the end-use sector most severely affected by the economic downturn in 2009. The growth rate for commercial energy use, at 1.1 % per year, is the fastest rate among the end-use sectors. US Commercial sector buildings must be targeted for improvement to make major gains in reducing US energy use.

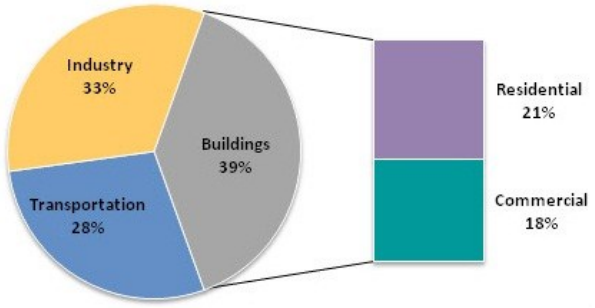


Figure 1: US energy consumption (US Department of Energy (DOE), 2008)

By 2035, approximately 75% of the built environment in the US is expected to be either new or renovated. Architecture 2030, a non-profit, non-partisan, independent organization has pointed out that this transformation represents a historic opportunity for the architecture and building community to reduce energy use, and thereby slow down climate change (Architecture 2030, 2010). This message has been crucial in spurring renewed interest in energy efficient building design and the various tools and processes associated with it.

Energy benchmarking offers an initial building energy performance assessment without rigorous evaluation. It is the process of comparing the energy performance of a particular commercial building to a range of energy-performance values of similar buildings, so as to rank the building in terms of energy efficiency among its peers, and then assess opportunities for energy efficiency. Just as Energy Guide labels on appliances indicate where the labeled appliance fits into the range of similar appliances from most to least efficient; benchmarking allows a ranking system for buildings to be defined.

Buildings are responsible for almost 40% of the greenhouse gas emissions (US EIA, 2008) and energy benchmarking is critical to improving building performance, thereby creating a healthy, green, and more livable environment (See Figure 2). There are many energy-related building codes, as well as various building-rating organizations that specify and rate the design of buildings. However, these design-based ratings are merely estimates, while benchmarking rates buildings based on measured energy consumption.

Some of the popularly used benchmarking tools are the Energy Star Portfolio Manager, ASHRAE building EQ, LBNL's Energy IQ which are described in detail in section 2.3. According to Energy Efficient Buildings Hub, benchmarking is a cycle of improvement. When buildings are provided with a rating, they tend to achieve market rewards for energy efficiency. The building owners continue to improve efficiency to stay competitive and therefore, the building efficiency keeps improving (See Figure 3).

The Clean and Affordable Energy Act of 2008 (CAEA) states that it is mandatory for owners of all large private buildings (over 50,000 gross square feet) in the district of Columbia to annually benchmark their energy and water efficiency and report the results for public disclosure (DDOE, 2013). Benchmarking is done using the United States Environmental Protection Agency's (US EPA) free, industry-standard online tool, ENERGY STAR ® Portfolio Manager. Final regulations of the act were published in January 2013. Energy disclosure laws in cities such as Austin, San Francisco, New York, Minneapolis, Philadelphia and Washington have made it compulsory for commercial buildings to be benchmarked for energy efficiency.

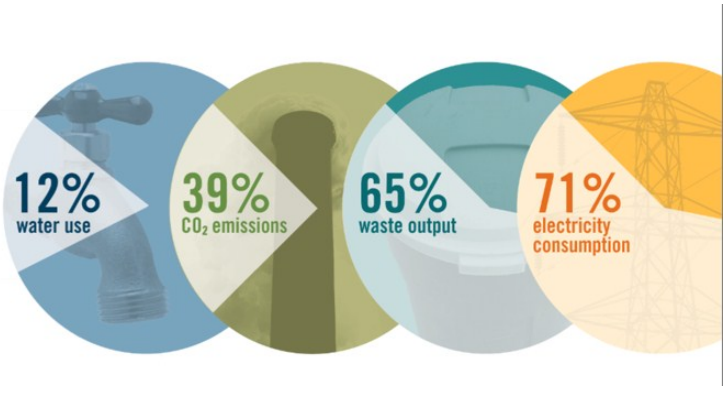


Figure 2: U.S. Building Impacts ("Green building alliance", 2013)

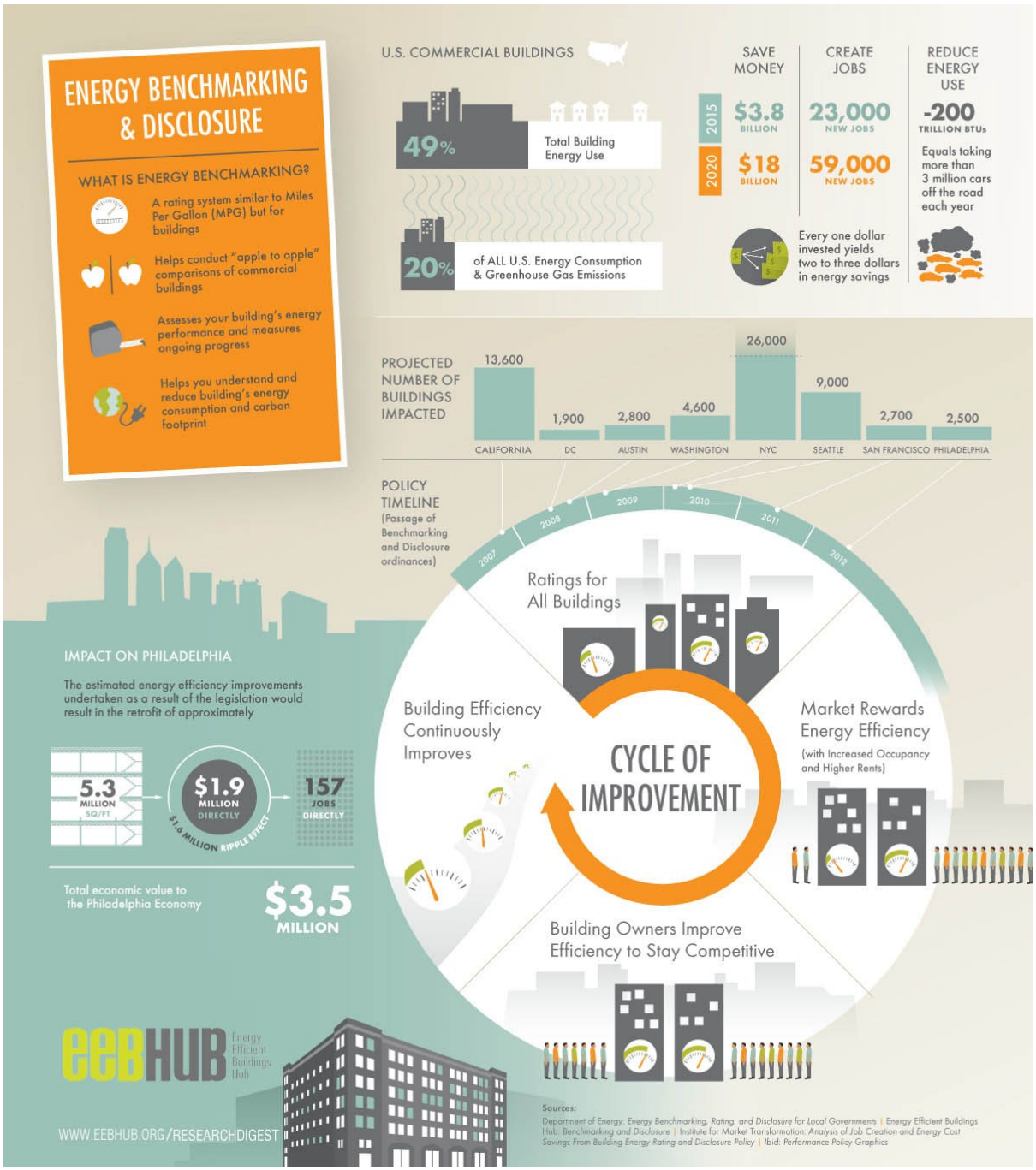


Figure 3. Cycle of Improvement (EEBHUB, 2013)

1.2 Problem Statement

An important issue in benchmarking is the use of performance indexes to characterize the building. The performance indexes sometimes serve as a benchmark by themselves. But, there are very few of them and not all are very reliable. The commonly used indexes are:

- *Comfort indexes*, comparing the actual comfort conditions to the comfort requirements
- *Energy indexes*, consisting of energy demand divided by heated/conditioned area, this allows comparison with reference values of the indexes coming from regulation or similar buildings
- *Energy demands* directly compared to "reference" energy demands generated from simulation tools
- *Energy Use Intensity (EUI)*, which is the rate of energy use (Energy consumption /conditioned area)

EUI is widely used as an energy benchmark in building energy analysis. It is expressed in kWh/sqft/yr or BTU/sqft./yr. EUIs are an attempt to normalize the energy use corresponding to a strong determinant (square footage) so that the energy use of many buildings is comparable. By normalizing out strong determinants, wide differences between building EUIs would be indicators of inefficient buildings or systems where improvements can be made (Sharp, 1996). EUIs are a standard unit of measurement for building energy analysis and have been studied for use as whole buildings energy targets. Despite being normalized for area, which is a strong determinant, EUIs vary considerably and are thus, ambiguous energy benchmarks as indicators of energy performance of an

individual building. To overcome this ambiguity, simple statistical models were developed to correct for variations in building characteristics. They were meant to be more accurate benchmarks or estimators of electricity use in a commercial building. Benchmarking tools that use this approach include: the U.S. Environmental Protection Agency's (EPA) ENERGY STAR® Portfolio Manager, Lawrence Berkeley National Laboratory's (LBNL) Energy IQ, ORNL's (Oak Ridge National Laboratory) spreadsheet and ASHRAE'S Building Energy Quotient.

The database that is widely used to obtain the entire country's commercial buildings energy information is the Commercial Buildings Energy Consumption Survey (CBECS) database. It contains energy consumption, energy expenditure and energy related building characteristics for 6,380 commercial buildings all over the US. Most of the variables in this database are categorical.

The most commonly used statistical method to develop these tools is the linear regression technique. **Linear regression** is a statistical method used to model a linear relationship between a scalar dependent variable '*y'* and one or more explanatory variables denoted '*X'*. The case of one explanatory variable is called *simple linear regression.* Linear regression is unlikely to yield optimal results, since non-linear relations cannot be captured and the technique is limited to continuous variables.

1.3 Objective

The aim of this research is to develop a methodology which allows one to quantify the importance of different variables that influence the EUI (electric use intensity) of a particular commercial building type, select the strongest variables and to develop a statistical modeling approach which can serve as a new benchmarking technique involving both categorical and continuous variables.

1.4 Scope

This research has been carried out to determine whether the CBECS data can be used to develop a new methodology based on data mining techniques that would be a dependable benchmarking model or estimator of electricity use of a particular commercial building type. This would provide a way to estimate electricity use for benchmarking an individual building to other similar buildings. Other modeling approaches proposed in the literature using the CBECS database were examined and compared with the new methodology. This study also allowed us to identify the dominant determinants of energy use of commercial buildings from the CBECS database. This methodology was applied only to two commercial building types namely: Office and School, however, it could be extended to other building types in the future.

CHAPTER 2

LITERATURE REVIEW

2.1 Base lining

A baseline is a point of reference from which comparisons are made. Measuring energy performance at a definite time establishes a baseline and provides the starting point for determining goals and evaluating future efforts and overall performance. Baselines should be established for all levels appropriate to any organization. (US EPA, 2013).

2.2 Benchmarking

A number of businesses are attempting to reduce their energy use by 30% or more through effective energy management practices. This involves gauging energy performance, setting energy savings goals, and regularly evaluating progress. And building-level energy performance benchmarking is integral to this process. It provides the reference points necessary for designing sound energy management practices and for gauging their effectiveness (US EPA, 2007).

Energy use benchmarking is a process that compares the energy use of a building or group of buildings with other similar structures. Alternatively, it may assess how energy use varies from a baseline. It is a critical step in any building upgrade project, since it helps organizations understand how and where they use energy and what factors drive their energy use. Further, it enables organizations to determine the key metrics for

assessing performance, establish baselines, and set goals. It also helps them identify building upgrade opportunities that can increase profitability by lowering energy and operating costs, and it facilitates continuous improvement by providing diagnostic measures to evaluate performance over time.

Benchmarking energy performance helps energy managers to identify best practices that can be replicated, either within a building or across a portfolio of buildings. Benchmarks can be reference points for measuring and rewarding good performance. They allow an organization to identify top-performing facilities for recognition and to prioritize poorly performing facilities for immediate improvement.

2.3 Benchmarking tools

There are many energy benchmarking tools in the market which are being used by architects, engineers and building owners. Some of the most popular tools are discussed below.

2.3.1 Energy Star Portfolio Manager

Portfolio Manager is an interactive energy management tool that allows tracking and assessing energy and water consumption across the entire portfolio of buildings in an online environment. This tool help businesses in setting investment priorities, identifying under-performing buildings, verifying efficiency improvements and receiving EPA recognition for superior energy performance (US EPA, n.d.).

Role of Energy Star Portfolio Manager

Once energy consumption and cost data is entered into the Portfolio Manager software, it enables the user to calculate building energy performance assess energy management goals over time, and identify strategic opportunities for savings. The tool allows streamlining of the portfolio's energy and water data through tracking key consumption, performance, and cost information portfolio-wide. The process involved in streamlining is as follows:

- Tracking multiple energy and water meters for each facility
- Customizing meter names and key information
- Benchmarking facilities relative to historical performance
- Monitoring percent improvement in weather-normalized source energy
- Monitoring energy and water expenditure
- Sharing building data inside or outside the organization
- Entering operating characteristics, tailored to each space-use category within the building

Technical Methodology

Portfolio Manager can provide EPA energy performance ratings for a range of building types. EPA's energy performance ratings are derived from U.S. energy and facility data. The ratings account for the impact of weather variations as well as physical and operating characteristics of each building. The energy performance of each building is rated on a scale of 1-100 relative to similar buildings nationwide. And buildings with superior

performance are eligible for EPA recognition. The ENERGY STAR label is awarded for facilities which fall in the top 25% of performance ratings nationally.

Portfolio Manager also calculates a building's greenhouse gas emissions from on-site fuel combustion and purchased electricity as well as district heating and cooling. While this is based on the amount of energy the building consumes, the emissions calculations have no bearing on the energy performance rating. The methodology for calculating greenhouse gas emissions is designed to be consistent with the Greenhouse Gas Protocol developed by the World Resources Institute and World Business Council for Sustainable Development (US EPA, n.d).

Energy consumption in buildings can vary up to 30% depending on local weather. Therefore, the Energy Performance Rating (EPR) removes the impact of weather by estimating the building's energy consumption for a "normal" weather year. Weather normalization is accomplished by performing a regression of one year of monthly energy consumption data against actual outdoor air temperatures. The 30-year average normal air temperature is then provided as input into the regression equation to determine the normalized energy consumption (Neida & Hicks, 2001).

The office building regression model is based on data from the Department of Energy, Energy Information Administration's 2003 Commercial Building Energy Consumption Survey (CBECS). The dependent variable in the office analysis is source energy use

intensity (source EUI) which is equal to the total source energy use of the facility per year divided by the gross floor area. The regression model analyses the key determinants of source EUI, i.e. those factors that explain the variation in source energy per square foot in offices.

The regression analysis identified the following six characteristics as the key explanatory variables for the expected average source EUI (kBtu/ft^2) in offices:

• Natural log of gross square foot of floor space

• Number of personal computers (PCs) per 1,000 square feet

• Natural log of weekly operating hours

• Natural log of the number of workers per 1,000 square feet

• Heating degree days * Percent of the building floor space that is heated

• Cooling degree days * Percent of the building floor space that is cooled

Each independent variable is centered relative to its mean value.

Table 1. Descriptive Statistics for variables in the final OLS model (US EPA, 2007)

Descriptive Statistics for Variables in Final Regression Model				
Variable	**Full Name**	**Mean**	**Minimum**	**Maximum**
SrcEUI	Source Energy per Square Foot	198.4	19.62	1133
LNSqFt	Natural Log of Square foot	9.535	8.517	13.82
PCDen	Number of Computers per 1000 ft2	2.231	0.0273	11.11
LNWkHrs	Natural Log of Weekly Operating Hours	3.972	3.611	5.124
LNWkrDen	Natural Log of Number of Workers per 1000 ft2	0.5616	-3.882	2.651
HDDxPH	Heating Degree Days x Percent Heated	4411	0.0000	9277
CDDxPC	Cooling Degree Days x Percent Cooled	1157	0.0000	5204

Example Calculation (US EPA, 2007)

The following is a specific example with the office model:

a) Step 1 – User enters building data into Portfolio Manager

(For the purposes of this example, sample data is provided)

• Energy data

- Total annual electricity = 3,500,000 kWh
- Total annual natural gas = 4,000 therms

(Note that this data is actually entered in monthly meter entries).

• Operational data

- Gross floor area = 200,000 ft^2
- Weekly operating hours = 80
- Workers on main shift = 250
- Number of personal computers = 250
- Percent heated = 100
- Percent cooled = 100
- HDD (provided by Portfolio Manager, based on zip code) = 4937
- CDD (provided by Portfolio Manager, based on zip code) = 1046

b) Step 2 – Portfolio Manager computes the Actual Source Energy Use Intensity. In order to compute actual source EUI, Portfolio Manager must convert each fuel from the specified units (e.g. kWh) into Site kBtu, and then from Site kBtu to Source kBtu.

• Convert the meter data entries into site kBtu

- Electricity: (3,500,000kWh)*(3.412kBtu/kWh) = 11,942,000 kBtu Site

- Natural gas: (4,000 therms)*(100kBtu/therm) = 400,000 kBtu Site

• Apply the source-site ratios to compute the source energy

- Electricity: 11,942,000 Site kBtu*(3.34 Source kBtu/Site kBtu) = 39,889,280 kBtu Source
- Natural Gas: 400,000 Site kBtu *(1.047 Source kBtu/Site kBtu) = 418,800 kBtu Source

• Combine source kBtu across all fuels

- 39,889,280 kBtu + 418,800 kBtu = 40,308,080 kBtu

• Divide total source energy by gross floor area

- Source EUI = 40,308,080 kBtu/200,000ft2 = 201.5 kBtu/ft^2

c) Step 3 – Portfolio Manager computes the Predicted Source Energy Intensity.

Portfolio Manager uses the building data entered under Step 1 to compute centered values for each operating parameter. These centered values are entered into the office regression equation to obtain a predicted source EUI.

• Calculate centered variables

- Use the operating characteristic values to compute each variable in the model. (e.g. LN(Square Foot) = LN(200,000) = 12.21)
- Subtract the reference centering value from calculated variable

(e.g. LN(Square Foot) - Reference centering value(see Table 2) =

LN(Square Foot) - 9.535 = 12.21 – 9.535 = 2.675).

(These calculations are summarized in Table 2)

• Compute predicted source energy use intensity

- Multiply each centered variable by the corresponding coefficient in the model (e.g. Coefficient*Centered LN(Square Foot) = 34.17*2.675 = 91.40)
- Take the sum of these products (i.e. coefficient*Centered Variable) and add to the constant (this yields a predicted Source EUI of 282.9 kBtu/ft^2)

(This calculation is summarized in Table 3).

Table 2. Computing Building Centered Variables (US EPA, 2007)

Example Calculation – Computing Building Centered Variables				
Operating Characteristic	Formula to Compute Variable	Building Variable Value	Reference Centering Value	Building Centered Variable (Variable Value - Center Value)
CLnSqFt	LN(Square Foot)	12.21	9.535	2.675
CPCDen	#Computers/ft^2*1000	1.250	2.231	-0.9810
CLNWkHrs	LN(Weekly Operating Hours)	4.382	3.972	0.4100
CLNWkrDen	LN(#Workers/ft^2*1000)	0.2230	0.5616	-0.3386
CHDDxPH	(HDD*Percent Heated)	4937	4411	526.0
CCDDxPC	(CDD*Percent Cooled)	1046	1157	-111.0
BANK_50xCLNSqFt	BANK_50*C_LNSqFt	0.0000	NA	0.0000
BANK_50xCLNWkrDen	BANK_50*C_LNWkeDen	0.0000	NA	0.0000
BANK_50	BANK_50	0.0000	NA	0.0000

Note
- *Densities are always expressed as the number per 1,000 square feet.*
- *The center reference values are the weighted mean values from the CBECS population, show in Table 2.*
- *Bank_50 has a value of 1 if the building is a bank of 50,000 square foot or smaller; otherwise it has a value of 0.*
- *The Bank_50 terms are not centered because they represent a multiplier on the already centered variables LNSqFt and LNWkrDen.*

d) Step 4 – Portfolio Manager computes the energy efficiency ratio

The energy efficiency ratio is equal to: Actual Source EUI/ Predicted Source EUI

• Ratio = 201.5/282.9 = 0.7123

Table 3.Computing Predicted Source EUI (US EPA, 2007)

Example Calculation – Computing predicted Source EUI			
Operating Characteristic	Centered Variable	Coefficient	Coefficient * Centered Variable
Constant	NA	186.6	186.6
CLnSqFt	2.675	34.17	91.40
CPCDen	-0.9810	17.28	-16.95
CLNWkHrs	0.4100	55.96	22.94
CLNWkrDen	-0.3386	10.34	-3.501
CHDDxPH	526.0	0.0077	4.050
CCDDxPC	-111.0	0.0144	-1.598
Bank_50xCLNSqFt	0.0000	-64.83	0.0000
Bank_50xCLNWkrDen	0.0000	34.20	0.0000
BANK_50	0.0000	56.30	0.0000
		Predicted Source EUI (kBtu/ft^2)	***282.9***

e) Step 5 – Portfolio Manager looks up the efficiency ratio in the lookup table

Starting at 100 and working down, Portfolio Manager searches the lookup table (Table 4) for the first ratio value that is larger than the computed ratio for the building.

• A ratio of 0.7123 is less than 0.7218 (requirement for 72) but greater than 0.7119 (requirement for 73)

• The rating is then chosen as **72**

When conducting regression analyses and when calculating energy performance ratings in Portfolio Manager, the actual reported energy use intensity and the actual HDD and CDD experienced by the building during the given timeframe are applied. Weather normalized source energy use intensity is not used in determining energy performance ratings (US EPA, 2011).

Table 4. List of energy efficiency ratio cut-off points for each rating, from 1 to 100 (US EPA, 2007)

Lookup Table for Office, Bank/Financial Institution, and Courthouse Rating							
Rating	Cumulative Percent	Energy Efficiency Ratio		Rating	Cumulative Percent	Energy Efficiency Ratio	
		>=	<			>=	<
100	0%	0	0.278705	50	50%	0.925442	0.935487
99	1%	0.278705	0.328379	49	51%	0.935487	0.945611
98	2%	0.328379	0.363070	48	52%	0.945611	0.955821
97	3%	0.363070	0.390860	47	53%	0.955821	0.966125
96	4%	0.390860	0.414570	46	54%	0.966125	0.976528
95	5%	0.414570	0.435548	45	55%	0.976528	0.987040
94	6%	0.435548	0.454556	44	56%	0.987040	0.997667
93	7%	0.454556	0.472069	43	57%	0.997667	1.008419
92	8%	0.472069	0.488407	42	58%	1.008419	1.019304
91	9%	0.488407	0.503796	41	59%	1.019304	1.030331
90	10%	0.503796	0.518402	40	60%	1.030331	1.041511
89	11%	0.518402	0.532352	39	61%	1.041511	1.052853
88	12%	0.532352	0.545744	38	62%	1.052853	1.064369
87	13%	0.545744	0.558657	37	63%	1.064369	1.076072
86	14%	0.558657	0.571154	36	64%	1.076072	1.087973
85	15%	0.571154	0.583289	35	65%	1.087973	1.100087
84	16%	0.583289	0.595105	34	66%	1.100087	1.112428
83	17%	0.595105	0.606640	33	67%	1.112428	1.125013
82	18%	0.606640	0.617925	32	68%	1.125013	1.137858
81	19%	0.617925	0.628989	31	69%	1.137858	1.150984
80	20%	0.628989	0.639856	30	70%	1.150984	1.164412
79	21%	0.639856	0.650546	29	71%	1.164412	1.178163
78	22%	0.650546	0.661079	28	72%	1.178163	1.192263
77	23%	0.661079	0.671471	27	73%	1.192263	1.206741
76	24%	0.671471	0.681738	26	74%	1.206741	1.221627
75	25%	0.681738	0.691894	25	75%	1.221627	1.236956
74	26%	0.691894	0.701950	24	76%	1.236956	1.252768
73	27%	0.701950	0.711919	23	77%	1.252768	1.269105
72	28%	0.711919	0.721810	22	78%	1.269105	1.286018
71	29%	0.721810	0.731635	21	79%	1.286018	1.303565
70	30%	0.731635	0.741401	20	80%	1.303565	1.321809
69	31%	0.741401	0.751118	19	81%	1.321809	1.340827
68	32%	0.751118	0.760793	18	82%	1.340827	1.360708
67	33%	0.760793	0.770434	17	83%	1.360708	1.381554
66	34%	0.770434	0.780049	16	84%	1.381554	1.403491
65	35%	0.780049	0.789645	15	85%	1.403491	1.426665
64	36%	0.789645	0.799227	14	86%	1.426665	1.451258
63	37%	0.799227	0.808804	13	87%	1.451258	1.477493
62	38%	0.808804	0.818380	12	88%	1.477493	1.505650
61	39%	0.818380	0.827963	11	89%	1.505650	1.536087
60	40%	0.827963	0.837558	10	90%	1.536087	1.569275
59	41%	0.837558	0.847171	9	91%	1.569275	1.605847
58	42%	0.847171	0.856808	8	92%	1.605847	1.646683
57	43%	0.856808	0.866475	7	93%	1.646683	1.693068
56	44%	0.866475	0.876178	6	94%	1.693068	1.746975
55	45%	0.876178	0.885923	5	95%	1.746975	1.811687
54	46%	0.885923	0.895716	4	96%	1.811687	1.893296
53	47%	0.895716	0.905563	3	97%	1.893296	2.005317
52	48%	0.905563	0.915469	2	98%	2.005317	2.190161
51	49%	0.915469	0.925442	1	99%	2.190161	>2.190161

2.3.2 ASHRAE Building Energy Quotient (BEQ)

ASHRAE Building Energy Quotient (BEQ) is a building energy labeling program which provides information on the potential and actual energy use of buildings. BEQ was introduced in 2009 as a pilot program with the intent of providing a simple scale to convey a building's energy use in comparison to similar buildings and climate zones. In addition, BEQ is also meant to provide building owners with building-specific information that highlights potential energy saving opportunities.

BEQ provides a range of benefits to various stakeholders (ASHRAE BEQ Program, 2009):

• Building owners and operators:

- Helps assess how their building compares against peer buildings, and establish a measure of their potential for energy performance improvement
- Allows differentiation from other buildings to attract potential buyers or tenants

• Potential buyers or tenants:

- Provides insight into the value and potential long-term cost of a building

• Operations and maintenance staff:

- Informs decisions on maintenance activities and helps influence building owners and managers to pursue equipment upgrades and demonstrate the return on investment for energy efficiency projects

New buildings are eligible to receive an asset rating (also called an "As Designed" rating). Buildings which have at least 12 months of consecutive energy use data are also eligible for an operational rating (also called an "In Operation" rating). The "As Designed" rating provides an assessment of the building based on a building's design specifications, for example, mechanical systems, building envelope, orientation, and daylighting. A field inspection and a building energy model are used to prepare the asset rating. The "In Operation" rating is prepared based on a combination of the structure of the building and how it is operated. Thus, it provides information on the actual energy use of a building. Information gained through successive years of operational labels can help building owners and operations and maintenance staff understand how the building performs, where opportunities for improvement lie, and where similar buildings fall in comparison. It also helps owners of portfolios of several buildings to identify priorities for energy savings investment (ASHRAE Building Energy Labeling Program Implementation Committee, 2009).

Table 5. Difference between Operational and Asset rating (ASHRAE Building Energy Labeling Program Implementation Committee, 2009)

Operational Rating - "In Operation"	Asset Rating - "As Designed"
• Objective is to improve operations • Rating based on measured energy usage, adjusted for weather • No inherent requirement for field verification • Ratings sometimes adjusted based on levels of service • Good for use in existing building energy efficiency incentive programs • Good for managing building portfolios over time • Example: U.S. EPA's ENERGY STAR® Portfolio Manager	• Objective is to value property • Rates the building, not the occupancy and operation. • Focus is on the physical building characteristics and permanent energy systems • Differences in operational behavior are ignored • Rating is derived from a modelbased estimate of energy usage, compared to a stock median or building code baseline for the building type • Field verification is a requirement • Good for evaluating building performance within a financial transaction • A basis for energy efficiency code compliance and beyond code new construction incentive programs. • Examples: RESNET and CEC Home Energy Rating Systems

The rating scales evaluating building energy performance are based on two general methods. *Statistical methods* use a frequency distribution of EUI of the population of buildings represented and provide a rating for a building according to its percentile location in the distribution (Figure 4). *Technical rating methods* compare a building's energy performance against technical potential reference points where Net Zero Energy performance is zero on the scale and the building type population median is set to 100 (Figure 5). The ASHRAE Building EQ is the same basic scale that is used in the European Union for commercial buildings and equivalent to the scale used in North America for the residential asset rating system (HERS - Home Energy Rating System). Comparisons of the two rating scales are shown in Table 6

Table 6. Difference between Statistical and Technical rating (ASHRAE Building Energy Labeling Program Implementation Committee, 2009)

Statistical Rating Scale	*Technical Rating Scale*
• Fit a regression model to a sample distribution of population data • Existing building population sample used to set low and high end of scale • Representative data required for the entire distribution of existing buildings of a particular type • Does not necessarily include energy policy goals in rating scale	• Rated buildings compared to stock median or code level of performance • Energy policy sets low end of scale (e.g. zero net energy or zero carbon) • Only stock median values are required for existing buildings of a particular type

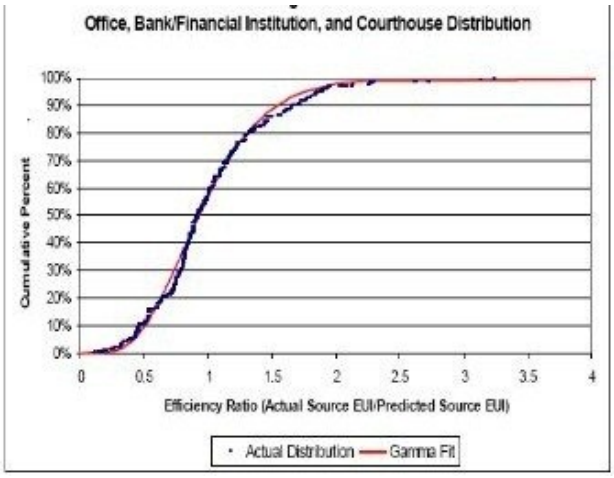


Figure 4. Statistical Rating Scale

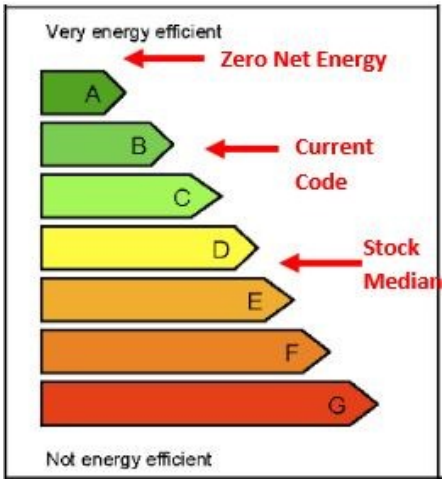


Figure 5. Technical Rating Scale

Figure 6 below shows an energy label given by BEQ.

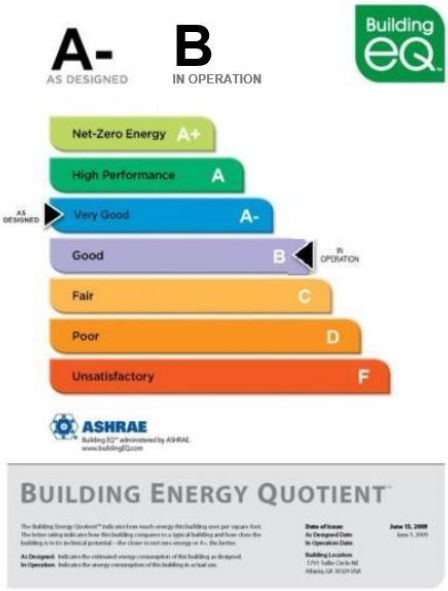


Figure 6. ASHRAE Building Energy Quotient Label

2.3.2 Energy IQ

The U.S. Department of Energy's Lawrence Berkeley National Laboratory has developed an energy benchmarking method that provides more practical guidance for energy efficiency improvement than traditional benchmarking tools. EnergyIQ is an "action-

oriented" benchmarking tool for non-residential buildings which provides a standardized opportunity assessment based on benchmarking results, along with decision-support information to help define action plans Action-oriented benchmarking helps identify options for energy efficiency improvements and prioritize areas for more detailed analysis. Such opportunity assessment is not afforded by conventional benchmarking tools. Action-oriented benchmarking improves on simplified benchmarking processes and lays the foundation for investment-grade audits and professional engineering calculations, as suggested in Figure7 (EnergyIQ, n.d.).

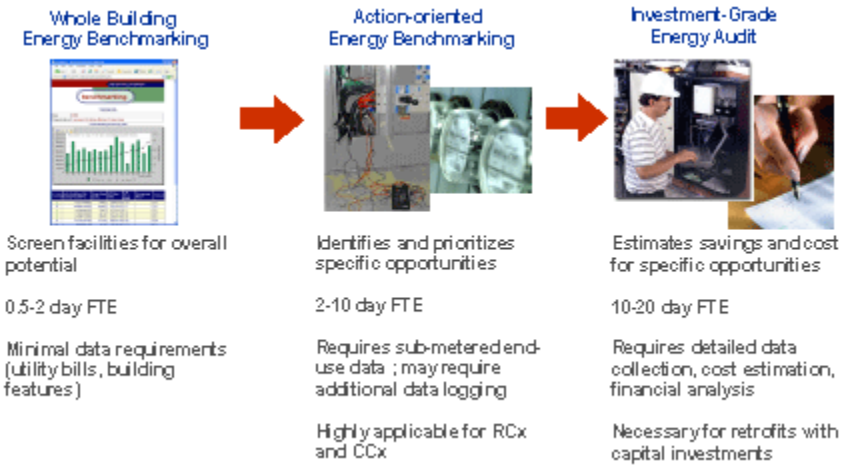


Figure 7. Action-Oriented Benchmarking (LBNL,2008)

EnergyIQ provides a more in-depth analysis compared to more generalized whole-buildings tools such as the ENERGY STAR Portfolio Manager. EnergyIQ benchmarks energy use, costs, and features for sixty two building types. In addition, it provides a carbon-emissions calculation for the energy consumed in the building (EnergyIQ, n.d.).

EnergyIQ has been designed to meet user needs which were identified through a survey carried out by LBNL as well as the outcomes of the ASHRAE Technical Research

Project-1286 best practices protocol for energy benchmarking tool design (Glazer, 2006). The tool includes multiple filters such as building type, location, vintage, floor area and size to enable the user to select an appropriate dataset. The user also has the option of evaluating portfolios of buildings individually or in aggregate. The tool provides access to a large database by accommodating the CBECS database in addition to CEUS database. The user has the option to include both databases as peer groups (as well as the results from other users of the tool) against which to compare a chosen building (Mills et al., 2008).

An important feature of the tool is that it minimizes the data required from the user by tailoring requirements to the desired output. To aid ease of use, EnergyIQ offers visual as well as tabular displays of benchmark metrics. Further, the tool supports benchmarking of a building to its peers at a single point in time, as well as benchmarking of the building to its own historical performance.

The tool generates a list of opportunities and recommendations based on user input. And the "Decision Support" module of the tool helps users to implement these recommendations by providing information on refining action plans and creating design-intent documentation (EnergyIQ, n.d.).

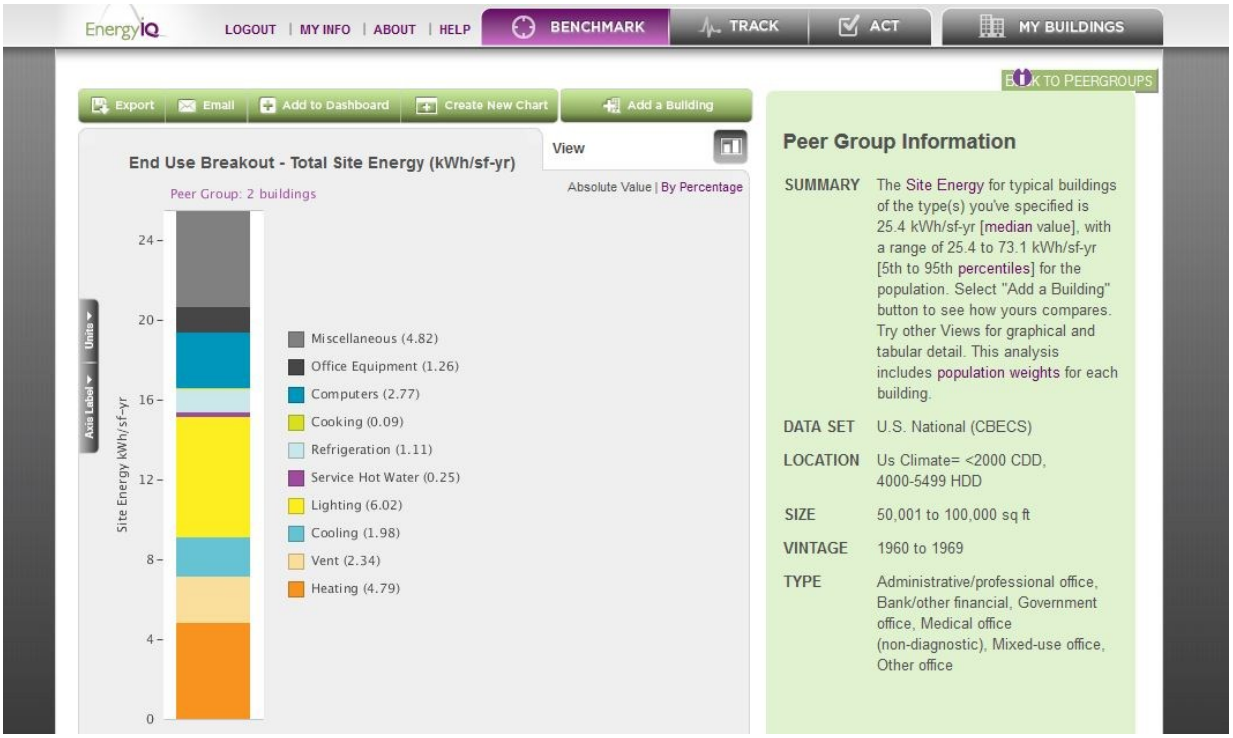


Figure 8. Energy IQ Result view

2.3.3 ORNL Spreadsheet

The Energy Use Intensity (EUI) distributions used in the spreadsheets developed by Oak Ridge National Laboratory (ORNL) were based upon a statistical analysis of approximately 1500 office buildings in the US Energy Information Administration's 1992 CBECS database. These were divided into their corresponding nine US census divisions for analysis. Thus, different areas of the US have different results depending on what characteristics were found most important to the locale. A subset of over 70 building characteristics from the CBECS database were selected and examined for their relationship to office building energy use. These were refined down to four characteristics that were the most important determinants of electricity use and the four

most important ones for nonelectric energy use. These few characteristics explained most of the variations in energy use that could be explained by considering all characteristics that had statistically significant relationships to energy use. Thus, addressing additional characteristics provided limited value. Within census divisions, climate was not a major driver of either electric or non-electric energy use (Sharp, 1996).

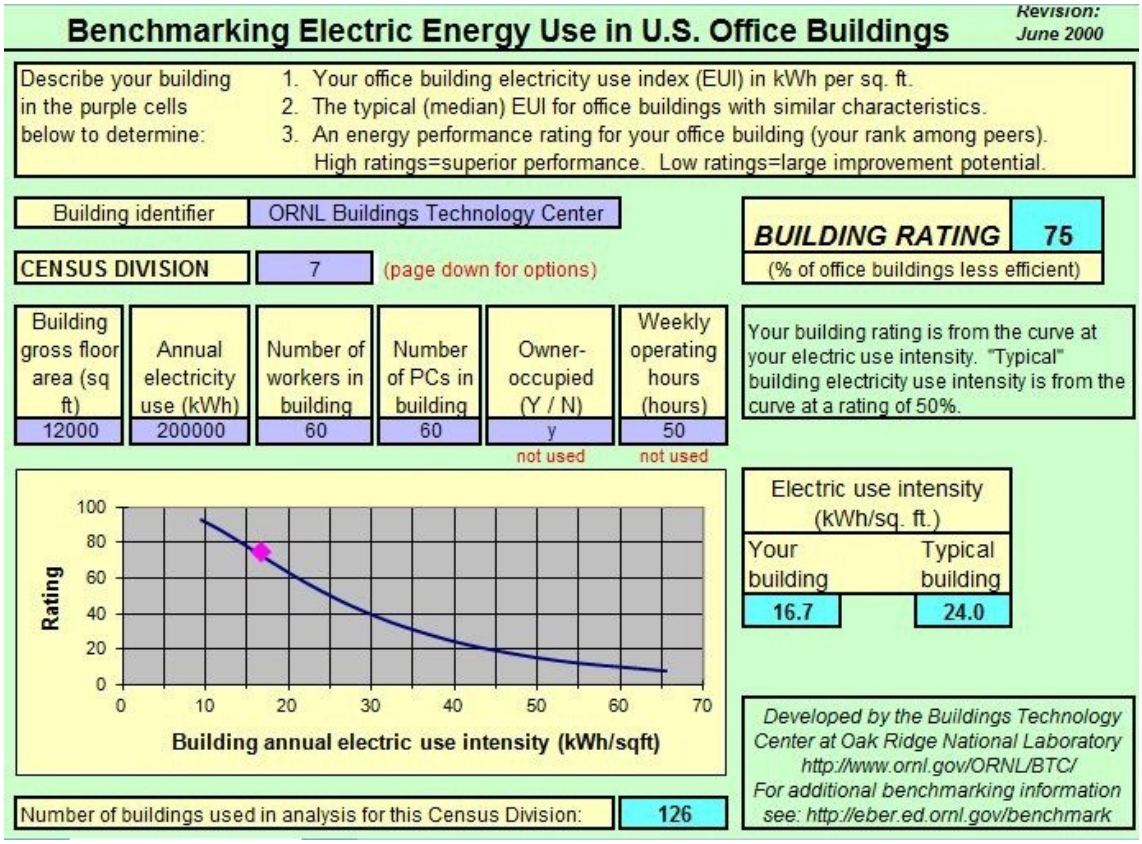


Figure 9. ORNL Spreadsheet view

The benchmarking spreadsheets developed by ORNL allow one to identify where one's specific office building ranks relative to others. They calculate the energy use intensity of the building, provide the median EUI for office buildings with the same characteristics, and identify where the building's performance ranks compared to others. They go beyond

the customary normalization by floor area and account for performance differences due to variations in worker density, the number of personal computers, operating hours, occupancy type, and heating fuel types. Beyond floor area, these characteristics were found to be the most common and most important drivers of electric and nonelectric energy use in US office buildings. Climate impacts on energy use were less significant, in part because analyses were conducted within regional census divisions (Sharp, 1996).

In this approach, the building is compared to others that have the same characteristics one provides as input. Thus, one is not comparing the building, which may have a high worker density (an important driver of energy use in 7 of 9 census divisions), to others with medium or low worker densities. Other important drivers of energy use can also be accounted for. Wide variances in these drivers can strongly impact the energy use in office buildings. By accounting for these, comparing office buildings that have sound reasons for higher energy use to those that do not is avoided. Average EUIs, although very commonly used, can be very misleading. This occurs because the distribution of energy use intensities for a group of buildings is normally highly skewed. This causes the average EUI for a group to be much higher than the median. For this situation, 65 to 70% or more of the buildings in many groups will often have lower EUIs than the group average. Many inefficient buildings will appear as moderate users in this situation. Small sample sizes can magnify this problem.

ORNL has also developed a brief table (see Table 7) which acts as an indicator of potential savings in the building after benchmarking the building using these spreadsheets. (ORNL, 1996).

Table 7. Rating and Energy Use indicators (ORNL,1996)

Rating for your building	Energy use and cost reduction potential (%)	Walk-thru energy assessment recommended?
below 20%	above 50%	Definitely
20 to 40%	35 to 50%	Yes
40 to 60%	20 to 35%	Maybe
above 60%	below 25%	No

Due to fuel cost differences and differing rate schedules, energy cost reduction percentiles should not be expected to exactly match energy use reduction percentiles. If a large portion of the energy costs consist of electric demand charges (often they make up 30-50% of a customer's electricity bill), the difference between energy use reduction percent and energy cost reduction percent can be significant.

2.3.4 FEDS (Retro-commissioning tool)

Facility Energy Decision System (FEDS)

The Facility Energy Decision System (FEDS) model is under development at the Pacific Northwest National Laboratory (PNNL) for the Department of Energy's (DOE) Federal Energy Management Program (DOE-FEMP), the U.S. Army Construction Engineering Research Laboratory (USA-CERL), the U.S. Army Forces Command (FORSCOM), the DOE's Rebuild America Program, the Defense Commissary Agency (DeCA), the U.S. Naval Facilities Engineering Service Center (NFESC), the Tennessee Army National

Guard, U.S. Army Installation Management Agency Southeast Region (IMA/SERO), U.S. Coast Guard (USCG) and Public Works and Government Services Canada (PNNL, 2008).

It is a user friendly building energy efficiency software tool used for assessing the energy efficiency potential of facilities ranging from single building to multi-building campuses and large federal installations. It identifies energy efficiency improvements quickly and objectively, that maximize life-cycle savings. The windows based, menu driven software requires only minimal user experience and input to perform energy efficiency assessment screenings as well as detailed energy retrofit project analyses (PNNL, 2011).

Some of the key features of the software are as follows:

• Requires minimal user input but also accepts detailed building system parameters. It approximates unspecified parameters based on typical characteristics for a building of the specified type, size, age, and location and other details.

• Simulates energy and cost performance of heating, cooling, ventilation, lighting, motors, plug loads, refrigeration, building shell, and hot water systems along with central plants and thermal loops.

• Computes energy consumption and fuel demand for each fuel type, technology, end use, building, and the entire installation.

• Provides a comprehensive approach to fuel-neutral, technology independent, integrated energy resource planning and acquisition.

•Assesses thousands of prospective energy efficiency options via a site optimized life-cycle cost minimization process.

• Reports investment requirements, net present value and payback period along with pre- and post-retrofit energy consumption and costs and air pollutant emissions impacts.

With minimal input, FEDS can be used as a top-down, first-pass energy systems analysis and energy resource acquisition decision software tool for buildings and facilities. Providing more detailed input allows the user to generate optimized building retrofits for an entire installation and provides detailed output for each retrofit in each building set. The basic intent of the model is to provide information needed to determine the minimum life-cycle cost (LCC) configuration of the installation's energy generation and consumption infrastructure. When determining the minimum LCC configuration of generation and end-use technologies, all interactive effects between energy systems are explicitly modeled. The value or cost of these interactive effects varies by building type (level of internal gain), building size (portion of heating, ventilation and air conditioning loads attributable to internal gains versus envelope gains/losses), climate (whether a particular building is cooling or heating-dominated), occupancy schedule and a number of other factors. Thus, there is no simple solution and detailed modeling, as is done in FEDS, is the best way to provide a credible estimate of the impact (PNNL, 2008).

The inferences about the building characteristics in FEDS are mostly obtained from the following sources:

• Non-residential Building Energy Consumption Survey (NBECS) and Residential Energy Consumption Survey (RECS) building characteristics data

• End-use Load and Conservation Assessment Program (ELCAP) commercial and residential end-use load and building characteristics data

• American Society for Heating Refrigeration and Air-conditioning Engineers (ASHRAE) standard design and construction practices.

The FEDS analysis process briefly consists of the followings steps:

1. Determine the building set breakdown

For large installations, with hundreds or thousands of buildings, FEDS is designed to model groups of buildings that can be categorized together into sets.

2. Complete an initial minimum set screening

A minimum set input provides a preliminary top-level screening indicating what actions should be initiated; further analysis is required before a project is designed and implemented.

3. Gather additional data about the buildings and central energy plants on the installation

Results from the minimum set screening are used to direct resources for additional data-gathering. The building types, end-uses and fuels with the largest potential savings (according to the screening) are the building types, end-uses and fuels that should be given the most time and money for additional data-gathering.

4. Select maximum detail display for selected building sets and modify inferred data

Maximum detailing in FEDS allows a knowledgeable user to override the default building and energy-using/generating equipment parameters that were inferred at

minimum set. Unlike other models that require detailed inputs, this approach allows but does not require the user to enter any site-specific information that is not readily available.

5. Set optimization parameters

The optimization parameters should be set to best suit one's needs. The following optimization parameter options should be taken into consideration:

• Select funding source

• Set financial screening options

• Exclude building sets that should not be considered for retrofits

• Restrict retrofit technologies or end uses that one does not want to evaluate

• Alter cost data

• Review emission factors

• Choose whether the output spreadsheet lists the optimal retrofits only or the top 3 retrofits

• Select any 'replacement required' flags for those technologies that must be replaced

6. Run model on final maximum detail input data

Once the data has been checked and modified by the user and inferred by FEDS, all building sets should be excluded from optimization before running FEDS to determine baseline consumption estimations. This allows the user to quickly get baseline information that is checked against real data and resolve any large discrepancies before doing a full run of the model. Once large discrepancies have been resolved, building set exclusions must be removed and the user should run the model.

2.4 Model used for energy benchmarking

There are various mathematical methods used in developing benchmarking systems. One of the methods is discussed in section 2.4.1.An approach to understand the importance of variation in building parameters is discussed in section 2.4.2.

2.4.1 Linear Regression Study

OLS (Ordinary Least Square) is the best known of all regression techniques. It is also the proper starting point for all spatial regression analyses. It provides a global model of the explanatory variables which helps in predicting an outcome. It generates a single regression equation to represent this model.

A study was conducted by Sharp, to identify the strongest determinants of office building energy use intensities. He found statistically significant relations between EUIs and several CBECS variables. The resulting performance model was used to predict the EUIs which were better benchmarks than simple census division statistics.

Seventy-five CBECS variables were selected to be examined as determinants of electric energy use intensity in office buildings. Stepwise regression was used to model electric energy use per square foot as a function of the CBECS variables. In the first analysis step 33 variables were found to be significant. Variables that were least significant and least common were removed in an iterative process. This produced six variables which were found to be the strongest determinants (Sharp, 1996).

Using these variables the predictive model for electric energy use intensity in commercial building is:

Log (kwhsf) = a +b*log (NWKERSF) + c*PCTRMC + d*OCCTYP1 + e*WKHRS + f*ECN + g*CHILLR (1)

where,

kwhsf = EUI (KWh/Sq.ft.)

NWKERSF= Number of workers per square feet

PCTRMC= number of personal computers

OOCTYP= occupancy type (yes/no categories)

WKHRS= Working hours

ECN= Economizer (yes/no categories)

CHILLR= Chiller (yes/no categories)

Sharp debated that mean EUI can be a poor benchmark because the distributions of indicators are usually skewed. So, he used the standard errors of the resulting regression model to establish the distributional benchmark table. This is considered more reliable as it masks the effect of outliers. The benchmarking process of a specific building makes use of the 'best-fitted' regression model to calculate the predicted EUI. With this predicted EUI, a distributional benchmark table is calculated through the mean values of the distribution of standard errors. The actual EUI can be compared with the table to get a score. A slightly modified version of this method is used by the Energy Star Portfolio Manager.

2.4.2 Simulation Study

Huang et. al (1993), studied the impact that variations in building conditions have on building's energy use patterns. They have used 481 prototypical commercial and multifamily buildings which were developed for the Gas Research Institute (GRI) to study the benefit of cogeneration for commercial buildings in 20 U.S. urban areas. These prototypical buildings for 13 U.S. cities, (defined by the authors in 1991) were then simulated using the DOE 2.1D program to create a database of the energy usage and hourly load shapes of those buildings. The study was conducted on two building types- large offices and hospitals for two locations- Chicago and Houston. Building characteristics such as the building size and the number of floors were obtained from a commercial company, F.W. Dodge, Inc. Building shell characteristics such as insulation levels, window areas and other information was derived from the CBECS database, 1989. Sensitivity analysis was used to develop a procedure to account for variations in building parameters to assess the market potential for specialized applications.

Parameters based on statistical sampling such as floor area and window percentage, the average values were increased and decreased by one standard deviation. For other parameters based on engineering judgment, the average values were modified up by 1.50 and down by 0.667.See Table 8 for the modified building parameters in the sensitivity analysis. (Huang et.al, 1993)

The sensitivity analysis for large offices indicated that their EUIs were highly influenced or sensitive to the lighting power density and hours of operation. Also, these EUIs were moderately sensitive to the building size and glazing characteristics for gas use and insensitive to the occupancy density, and building size and glazing characteristics for electricity usage. (Huang et.al, 1993)

Table 8. Altered building parameters in sensitivity analysis (Huang et.al, 1993)

Building Parameter	***Increase***	***Decrease***
Building size	+1 Standard Deviation	-1Standard Deviation
Glazing Percentage	+1 Standard Deviation	-1Standard Deviation
Insulation	ASHRAE 90.1	None
Number of Occupants	+1 Standard Deviation	-1Standard Deviation
Hours of Operation	24 hours/day	8 hours/day
Lighting power density	1.5 average	0.667 average
Equipment power density	1.5 average	0.667 average
Extreme bounding conditions-High/Low EUI	Building with small size, large glazing area, high occupancy, long hours of operation & high lighting and equipment power density	Building with large size, small glazing area, low occupancy, short hours of operation & low lighting and equipment power density

This research helped in understanding that the it is important to define the range and variability of end-use conditions in commercial buildings rather than knowing their shell conditions. Most of the simulation efforts, the building is most precisely defined in terms

of its physical attributes, lesser for equipment and operating schedules and almost random for end-use intensities.

The next chapter describes the methodology carried out for this research.

CHAPTER 3

METHODOLOGY

The overall benchmarking methodology proposed in this study is given in the flowchart in Figure 10. The first step is to select a building type and location closest to the building being evaluated. For this study, office and school buildings were selected. The relevant data from CBECS public use data 2003 was then extracted and filtered. Then, the outliers were flagged and removed from the data to reduce the variability in the data. Some of the variables were normalized to adjust the data. Out of all the variables in the database, the most significant variables were selected. In the next step the continuous variables were discretized (converted to categorical variables) to ensure that all the variables are uniform. Then, random forest method (regression and classification version) was used to determine the important variables. Using these variables, a single regression tree was generated for conceptual interpretability.

An additional step in the analysis was to compare the OLS model used by the Portfolio manager with the linear regression model designed with the selected data. The variables used in the linear regression model used by the Portfolio Manager ware used for this data. Sharp's model (Sharp, 1996) was also run using the office data used for this research. This was done to assess the data with existing models.

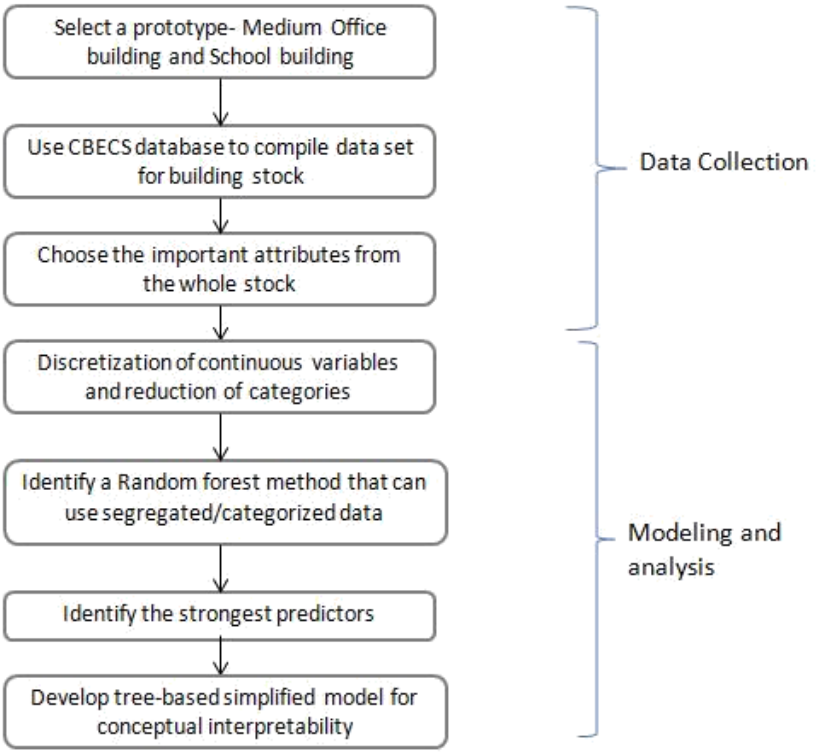


Figure 10. Baseline Methodology

3.1 Selecting bldg type and variables

The CBECS database consists of weighting factors to weight each building in proportion to the number of buildings of the same type in the U.S.These weighting factors were not used for this experiment. Therefore, each building in the database represents a single building.

This was done to keep the analysis simple. Also, the ambiguity that these specific weighting factors would bring appropriate representations as individual building characteristics like those resulting from this analysis can vary a lot from building to building. This methodology uses 244 out of about 1450 office buildings and 223 school

buildings from the CBECS data set. Buildings were selected by using the following filters:

a) Building Type:

   Among all the commercial building types in U.S, Office buildings consume the maximum energy. School buildings follow the same trend of energy consumption as office buildings. Therefore, these building types were selected.

b) Area Filter:

   The buildings were filtered by area. Medium-sized (15,000- 60,000 sq.ft.) office buildings were selected. The CBECS database provides a weighted average for buildings greater than 1,000,000 square feet. These buildings were removed to avoid biasing in the EUI results.

c) Outliers: The data sets were filtered for outliers or data points that were unusually high or low. See section 3.3.

3.2 Normalizing the variables

The intention of normalizing or adjusting the data is to bring the entire probability distributions of adjusted values into alignment i.e., within a certain specified range. This eliminates the effect of certain gross influences (e.g.: Square footage). See Section 4.1 and 4.2.

3.3 Removing Outliers

An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs. The circled points in Figure 11 represent the outliers.

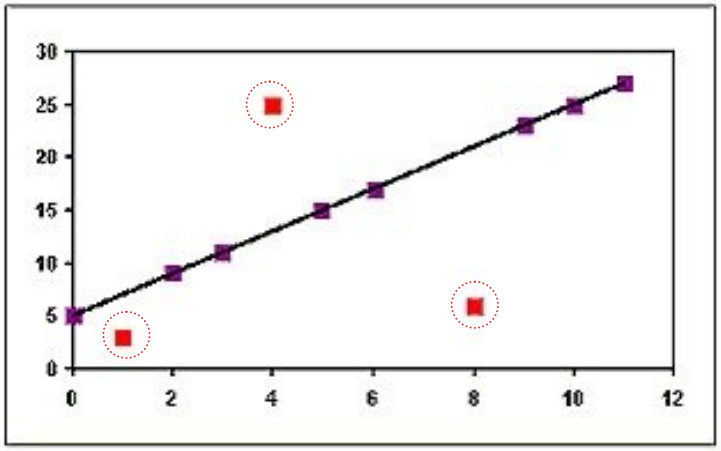


Figure 11. Example of Outliers ("Mark Young Training Systems", n.d.)

The method used to flag observations was based on the inter-quartile range. If Q_1 and Q_3 are the lower and upper quartiles respectively, then an outlier is defined to be any observation outside the range:

Criteria to flag outlier: $[Q_1 - k(Q_3 - Q_1), Q_3 + k(Q_3 - Q_1)]$ (2)

for some non-negative constant k, which is selected by the user. It can range from +3 to -3 so that it lies within the normal distribution. Larger the value of k, larger is the top and bottom threshold for the outliers ("Identifying outliers", 2013).

The **inter-quartile range** (**IQR**), also called the **mid-spread** or **middle fifty**, is a measure of statistical dispersion, being equal to the difference between the upper and lower quartiles ("Inter-quartile range", 2013),

$$\text{IQR} = Q_3 - Q_1 \quad (3)$$

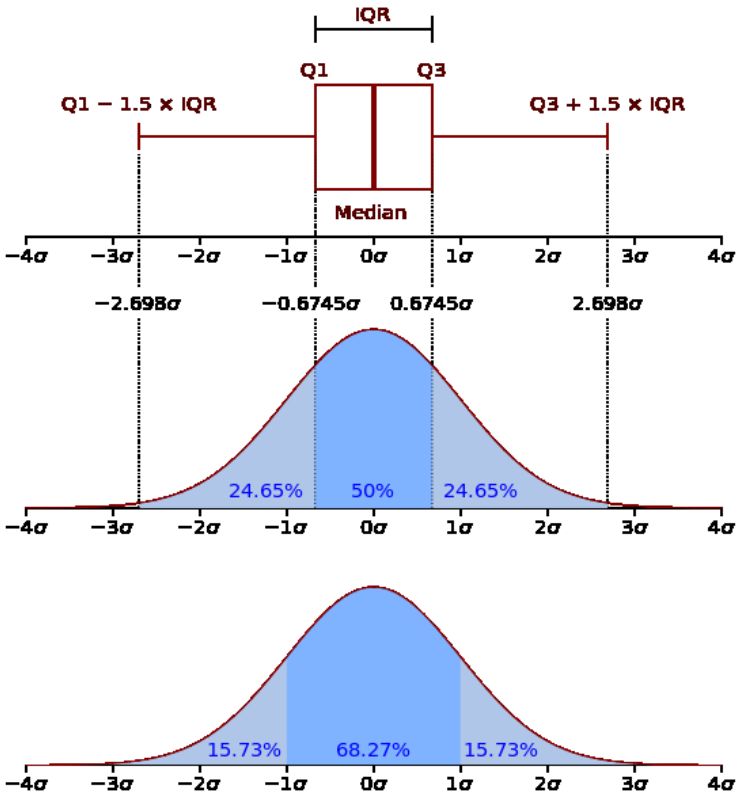


Figure 12. Box-plot and probability density function of a normal distribution ("Inter-quartile range", 2013)

The outliers lying in the region greater than 2σ (or 2 standard deviations) were removed.

3.4 Calculating the variability in EUI

The variability in EUI can be calculated using mean squared error or coefficient of variation. See Section 3.5.

3.5 Linear Regression

Regression analysis allows one to model, examine, and explore spatial relationships, and can help explain the factors behind observed spatial patterns. Regression analysis is also used for prediction. In statistics, **linear regression** is a method to model the relationship

between a scalar dependent variable *y* and one or more explanatory variables denoted *X* ("Linear Regression", 2013).When it is one explanatory variable it is called *simple linear regression*. When it is more than one explanatory variable it is called *multiple linear regression*. Equation (4) describes a simple linear regression.

$$\mathbf{y= mX+c} \qquad (4)$$

where, y is the dependent variable, X is the independent variable, m is the slope and c is the constant.

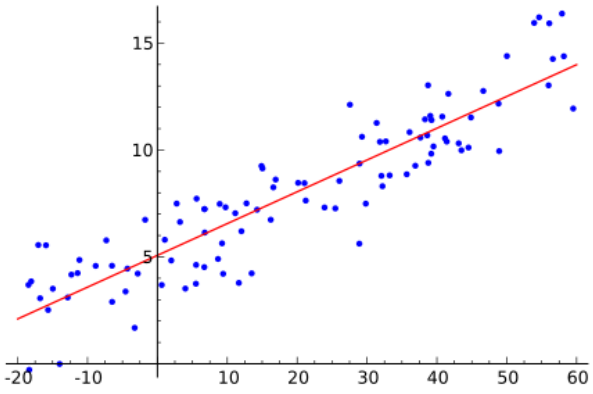


Figure 13. Simple linear regression (Wikipedia)

Sum of Squares

Mathematically, the sum of squared deviations is an unadjusted measure of variability. The distance from any point in a dataset, to the mean of the data, is the deviation. This deviation can be written as $y_i - \overline{y}$, where y_i is the ith data point, and $\overline{y}$ is the estimate of the mean. If all such deviations are squared, then summed, then

$$\text{Sum of Squares} = \sum_{i=1}^{n} (y_i - \overline{y})^2, \qquad (5)$$

("Partition of Sum of Squares", 2013)

Ordinary least squares (OLS) is the simplest method of linear regression and thus most commonly used estimator. It is conceptually simple and straightforward in computation. OLS estimates are commonly used to analyze both experimental and observational data. The estimator is unbiased and consistent if the errors have finite variance and are uncorrelated with the regressors.

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \hat{e} \quad (6)$$

where, y is the dependent variable, β's are the coefficients. ("Mistakes to avoid and reporting OLS", n.d.)

The residual, ê, is the difference between the actual 'y' and the predicted 'y' and has a mean which is zero. It means that, OLS calculates the slope coefficients so that the difference between the predicted 'y' and the actual 'y' is minimized. The residuals are squared so that negative errors can be easily compared to positive errors. ("Mistakes to avoid and reporting OLS", n.d.)

Mean Squared Error (MSE)

For an unbiased estimator, the MSE is the variance of the estimator. If $\hat{y}_t$ represents the predictions of the regression's dependent variable, and y represents the true values of this variable, then the (estimated) MSE of the predictor is ("Mean squared error", 2013):

$$\text{MSE} = \frac{\sum_{t=1}^{n} (y_t - \hat{y}_t)^2}{n} \quad (7)$$

MSE helps in evaluating the variability and bias of predictions.

Root Mean Squared Error (RMSE)

The **root-mean-squared error (RMSE)** is another commonly used measure of the differences between values predicted by a model and the values actually observed. The RMSE of predicted values $\hat{y}_t$ for times *t* of a regression's dependent variable y is computed for *n* different predictions as the square root of the mean of the squares of the deviations ("Root-mean-square deviation", 2013):

$$\text{RMSE} \quad \overline{\frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{n}} \tag{8}$$

Coefficient of Variation (CV)

The coefficient of variation (CV) is the ratio of the standard deviation σ (or the RMSE) to the mean μ:

$$\text{CV} = \frac{\sigma}{\mu} = \frac{\text{RMSE}}{\mu} \tag{9}$$

A lower CV indicates that there is lower variation in the distribution of the data. It shows the extent of variability in relation to mean of a population. ("Coefficient of variation", 2013). The CV is a dimensionless number and is useful for comparisons between data sets with different units or widely different means.

3.6 Data mining approach

Predictive analytics encompasses a variety of techniques from *statistics, modeling* and *data mining* that analyze given facts to make predictions about the unknown. (Nyce,

2007). **Predictive models** analyze previous performance to assess how likely a building is to exhibit a specific behavior in order to improve efficiency.

Data mining is an interdisciplinary sub-discipline of computer science. It is the computational process of identifying patterns and relationships in large data sets. This involves a combination of tools from artificial intelligence, machine learning, statistics, and database systems (Clifton, 2010).

The overall goal of the data mining process is to extract information from a data set and transform it into a comprehensible format for further use.

3.6.1 Decision tree

Decision tree is a predictive tree-like model in which the inner nodes represent the test on an attribute, each branch represents the result of test and each leaf node represents the decision taken after computing all attributes ("Decision Tree", 2013). The final result is called the Terminal node. A path from root to leaf represents classification rules (See Figure 14).

In decision analysis a decision tree is used as a visual and analytical decision support tool, where the expected values of competing alternatives are calculated. Decision trees are commonly used in research, specifically in decision analysis, to help identify a strategy most likely to reach a goal ("Decision Tree", 2013). This study is an effort to use decision tree technique for energy benchmarking.

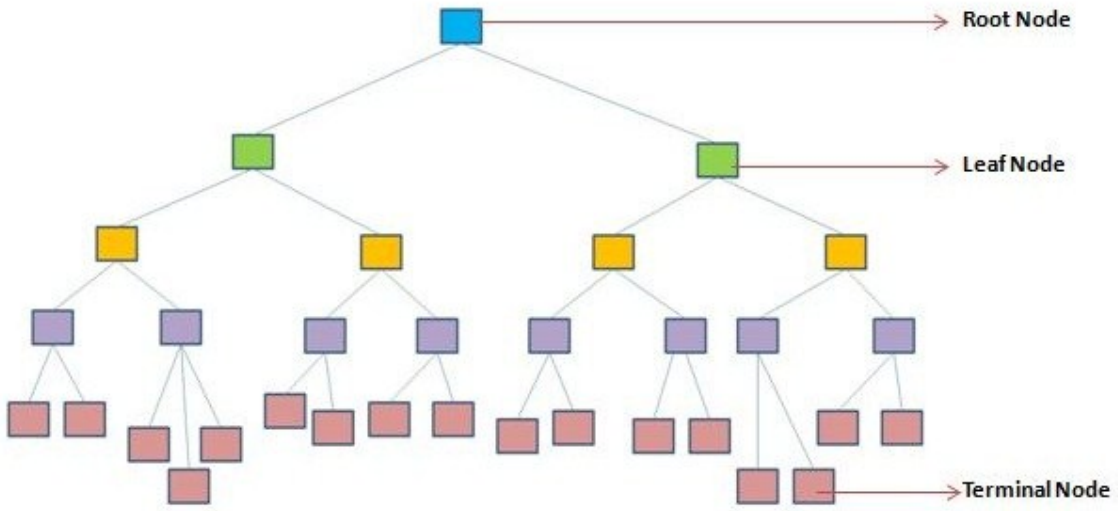


Figure 14. Decision tree format

Recursive partitioning or hierarchical clustering is a statistical method for multivariable analysis. Recursive partitioning generates a decision tree that strives to correctly classify members of the population based on several dichotomous dependent variables (Breiman et.al, 1984). As compared to regression analysis, which creates a formula that building owners can use to calculate the probability of energy use, recursive partition creates a rule such as 'If a building has variables x, y, or z the energy use is probably q'.

Figure 15 shows how this approach partitions or sub-divides the space into smaller regions, where the interactions are more manageable. The partitioning continues until the sub-divisions are so complaint that a simple model can be fit into them. The global model has two parts: recursive partition and the other is a simple model for each cell of the partition. The tree represents this process. Each terminal node or leaf represents a cell of

the partition and attached to it is a simple model that applies to that cell only (Breiman et. al, 1984).

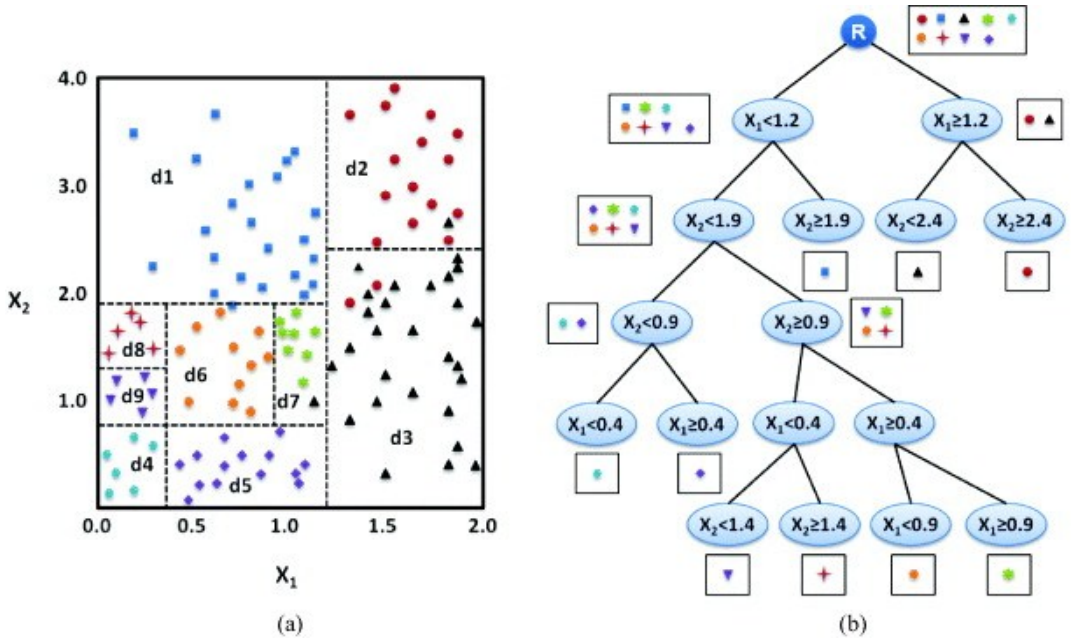


Figure 15. Recursive Partitioning (Kong et al., 2012)

Decision trees used in data mining are of two main types ("Classification and regression trees", 2009):

- **Regression tree** analysis is when the predicted outcome can be considered a real number or continuous.
- **Classification tree** analysis is when the predicted outcome is the class or a discreet category rather than a numeric value to which the data belongs. The variables which go into the classification (the inputs) can be numerical or categorical themselves, the same way they can with a regression tree. Like

regression trees, they provide moderately comprehensible predictors in situations where there are many variables which interact in complicated, nonlinear ways.

3.6.2 Decision tree Algorithm

For Regression Tree

In statistics, the **mean squared error (MSE)** of an estimator is one of many ways to quantify the difference between predicted and the true values of the quantity being estimated (See Equation 7 for the equation to calculate MSE). MSE assesses the quality of a set of predictions in terms of its variation and degree of bias ("Classification and regression trees", 2009).

For Classification Tree

A diversity index is a quantitative measure that indicates the number of different types that are present in a dataset, along with that it concurrently takes into account how evenly the basic entities are distributed among those types. The value of a diversity index increases both when the number of types increases and when evenness increases ("Diversity Index", 2013).

Gini's Diversity Index (gdi) — The Gini index of a node is

$$1 - \sum_i p^2(i), \tag{11}$$

where the sum is over the classes *i* at the node, and *p*(*i*) is the observed fraction of classes with class *i* that reach the node. A node with just one class (a *pure* node) has Gini index 0; otherwise the Gini index is positive. So this index is a measure of node impurity ("Gini's Diversity Index", n.d.).

3.6.3 CART Performance Metrics

Measures of Fit

Cross-validation is mainly used to determine the anticipated level of fit of a model to a dataset that is independent of the data used to train the model. It can also be used to determine quantitative measure of fit that is appropriate for the data and model. For example, for classification problems in each case the prediction is a unique class, in such a situation, **misclassification error rate** is used to summarize the fit. When the value predicted is continuous or numeric then, the **mean squared error**, **root mean squared error** or **median absolute deviation** are used to summarize the errors or fit ("Cross validation", 2013).

Cross validation

A decision tree starts with a single node, and then scans for the binary distinction which gives most information about the class (See Recursive Partitioning explained early in section 3.6.1). Next, it takes each of the resulting new nodes and repeats the process, continuing the recursion until it reaches some stopping criterion. The resulting tree is often too large (i.e., over-fit), so it can be pruned or cut back using cross-validation.

Cross Validation is also known as **rotation estimation**. It is a model validation technique for assessing whether the results of an analysis would generalize into an independent data set. It is used when the goal is a prediction, and to evaluate the accuracy of a predictive model. Single round of cross-validation involves partitioning a sample of data into equivalent subsets, performing the analysis on one subset (which is known as the *training set)*, and validating the analysis on the other subset (which is known as the *validation set* or *testing set*). To reduce inconsistency, multiple rounds of cross-validation are performed using different partitions, and the validation results are averaged over all the rounds (Kohavi et al., 1995).

Misclassification Error Rate

One of the common ways of measuring error is misclassification rate. It is the fraction of cases assigned in wrong classes. This is an index which is typical for classification trees only. A confusion matrix helps in determining the misclassification rate of classes. A confusion matrix helps in determining the misclassification rate of classes. Confusion matrix is table layout of actual class (rows) and predicted class (columns). The diagonal of this table consists of the correct classification or prediction. The other cells show the number of misclassifications. This table is useful for visually inspecting the errors.

Variable Importance

The variable importance is determined by the reduction in MSE. Lesser the value of MSE, greater is the variable importance.

3.6.4 Random Forest

Random forest is an ensemble learning method for classification (and regression) that operates by constructing a collection of decision trees at training time and outputting the class that is the mode of the classes output by individual trees. The algorithm for inducing a random forest was developed by Leo Breiman and Adele Cutler, and "Random Forests" is their trademark (Breiman, 2001). The term came from **random decision forest** that was first proposed by Tin Kam Ho of Bell Labs in 1995. The method combines Breiman's "bagging" idea and the random selection of features, introduced independently by Ho, Amit and Geman in order to construct a collection of decision trees with controlled variation (Ho et. al, 1995)

Random Forest Algorithm

The random forest algorithm has excellent accuracy and gives good insights into the inside of the box. It is a black box model since there is no interpretability to this model. There is classification version and a regression version.

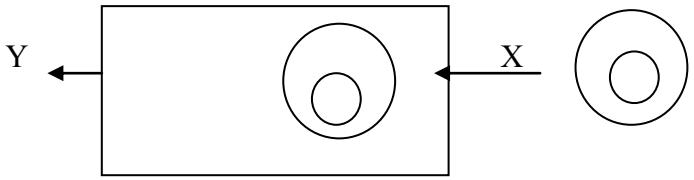


Figure 16. Black Box Model (Breiman, 2003)

a) Right eye in the model

The right eye in the model is called the classification machine (See Figure 16). This part of the model gives excellent accuracy and is an internal unbiased

estimate of test set error as trees are added to the ensemble. It can handle thousands of variables, many valued categorical, extensive missing values, badly unbalanced datasets and it cannot over-fit (Breiman, 2003).

b) Left eye in the model

This part of the model is inside the black box. It provides variable importance and outlier detection (Breiman, 2003).

c) Out of bag error

For every tree grown, about one-third of the cases are out-of-bag (oob). The oob samples serve as a test set for the tree grown on the non-oob data. This is used to form unbiased estimates of the forest test set error as the trees are added and forms estimates of variable importance (Breiman, 2003).

3.6.5 Advantages and Disadvantages of Decision trees

Advantages of using Decision Trees ("Decision Tree learning", 2013)**:**

- **Uncomplicated and easy to interpret.** It is a very intuitive method. It is easy to understand decision tree models after a brief explanation.
- **Minimum data preparation.** Other techniques often require normalization of data, creation of dummy variables and removal of blank values.
- **Able to handle both continuous and categorical data.** Other techniques are usually specialized in analyzing datasets that have only one variable type – usually continuous, whereas, a decision tree can handle both the variable types.

- **Uses a transparent model.** If a situation is perceptible in a model the explanation for the condition is easily explained by Boolean logic. Decision trees are easy to comprehend because of their very perceptible model.
- **Possible to validate a model using statistical tests.** These tests account for reliability of the model.
- **Robust model.** Decision trees perform well even if its assumptions are a bit violated by the true model from which the data were generated.
- **Performs well with large datasets.** Decision trees allow large amounts of data can be analyzed using standard computing resources in reasonable time.

Disadvantages of using Decision trees ("Decision Tree learning", 2013)**:**

- Decision-tree learners tend to create complex trees that do not generalize well from the training data. (This is known as over fitting). Mechanisms such as pruning are necessary to avoid this problem.
- Some concepts are difficult to learn because decision trees do not express them well.
- For data including categorical attributes with different numbers of categories, information gain in decision trees is biased in favor of those attributes with larger number of categories.

3.7 Reducing the categories in the variables/ Discretization

The variable selection technique should provide a measure of importance for each variable, instead of just listing them. This would provide more information for modifying the selected list to accommodate other considerations of the data analysis. Intuitively, the variables used in a tree have different levels of importance.

The CBECS database has provided most of the data as categorical variables. Each variable (e.g.: main cooling equipment, number of floors etc) has different number of categories. They range from two classes to eleven classes. With such a diverse number of classes the probability of acquiring a greater misclassification error rate is high. Therefore, the classes for each of the 18 variables, was reduced to three or four classes.

This was done in two ways:

(a). Dividing the data equally into each class for the respective variable

(b). Grouping similar categories or dividing the normal distribution

See Section 4.2 and Appendix B & Appendix C for further discussion on how the two divisions were done.

CHAPTER 4

APPLICATION TO CBECS DATA

4.1 Commercial Buildings Energy Consumption Survey (CBECS)

The 2003 Commercial Buildings Energy Consumption Survey (CBECS) database contains building characteristics, system descriptions, energy expenditure, and energy consumption for 6,380 commercial buildings across the US. Commercial buildings include office buildings, schools, correctional institutions and buildings used for religious worship.

This data represents all the fifty states and the District of Columbia (See Figure 17). There are about 1400 office buildings in this database.

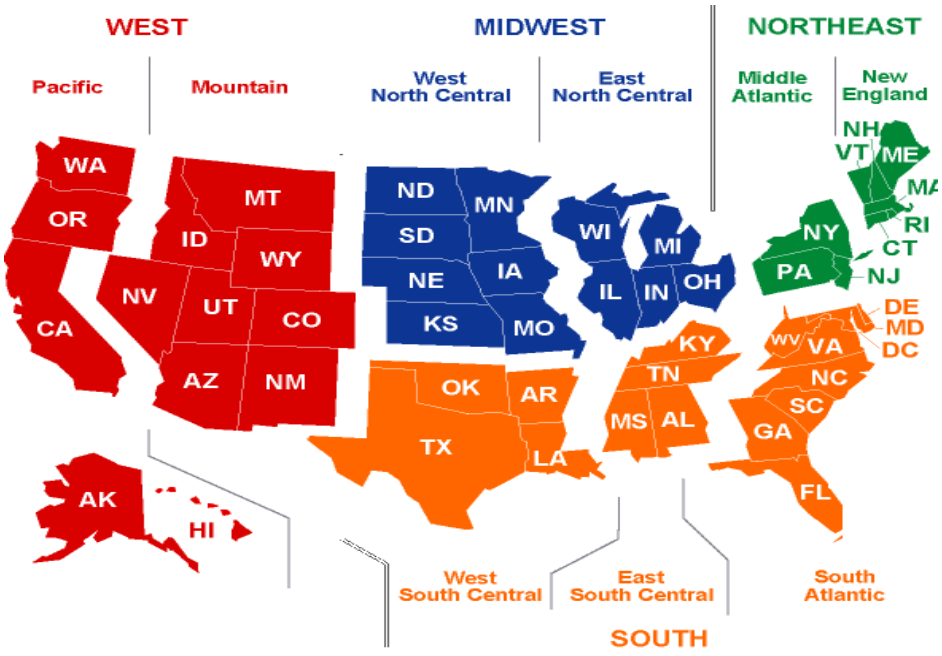


Figure 17. Region and Census division Map (CBECS, 2003)

All the buildings have over 1000 square feet of floor area. The reported floor areas have been rounded within square footage categories, except for buildings greater than one million square feet. As a result, errors can occur in the reported floor areas.

The National Oceanic and Atmospheric Administration (NOAA) have defined the climate zones as groups of climate divisions, which are regions within a state that are as climatically consistent as possible. Each climate division is placed into one of the five CBECS climate zones based on its 30-year average cooling degree-days (CDD) and heating degree-days (HDD) for the time frame between 1971 through 2000. (These climate zones have been updated for the 2003 CBECS. The previous database used averages for the 45-year period from 1931 through 1975). See Figure 18 for the climate zones.

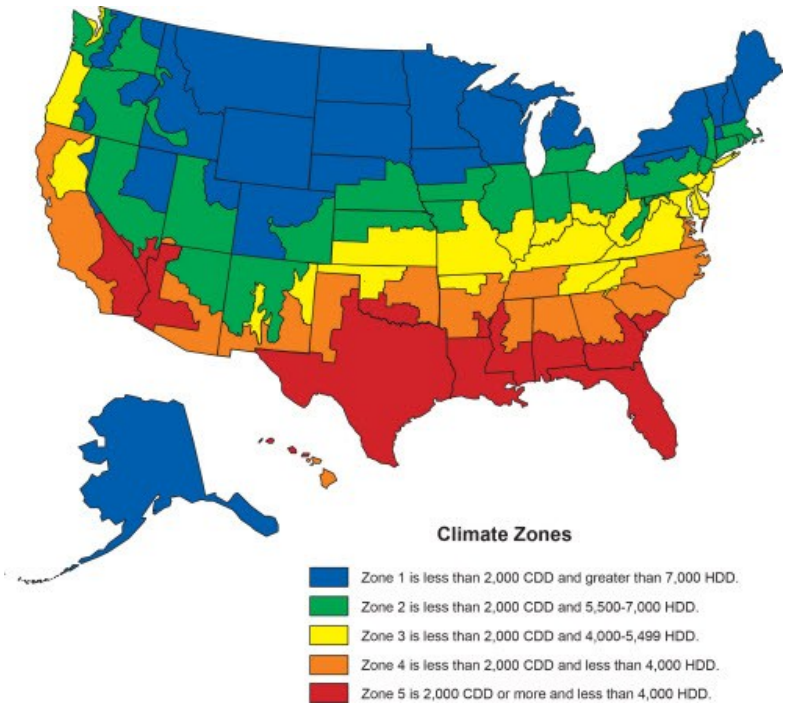


Figure 18. Climate Zones (CBECS, 2003)

The method used for data collection was Computer Assisted Personal Interviews with building owners, tenants and managers. Consequently, their accuracy depends on how well the respondents knew about their buildings. Sharp (1996) found that when reported floor areas for square and rectangular buildings were compared to their calculated floor areas (which were calculated based on number of floors, building length and width at ground level reported by building managers, owners and tenants in the personal interviews) large discrepancies were found. Figure 19 shows that many calculated floor areas are much smaller than the reported floor areas. Usually the calculated areas must match or exceed the reported areas. Calculated areas that contradict this appear for many buildings. Errors in length and width data were most suspected due to strong correlations between reported floor areas and electricity use.

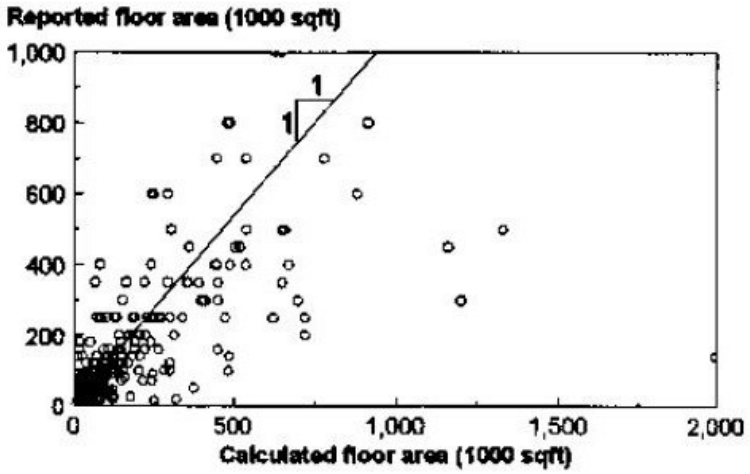


Figure 19. Electricity Use vs. Reported Floor area (Sharp, 1996)

The first survey was conducted in 1979. The most recent survey will be launched early April 2013 to provide data for the calendar year 2012. CBECS is currently updated on a quadrennial basis (US EIA, 2013).

4.1.1 Office

Medium-sized office buildings were selected for this research. The floor areas of the selected buildings were in the 15,000 – 60,000 sq.ft range. Based on the afore-mentioned range, 242 buildings were selected.

Bin	*Frequency*
5000	0
10000	89
15000	57
20000	61
25000	27
30000	31
35000	30
40000	19
45000	25
50000	18
55000	15
60000	8
More	0

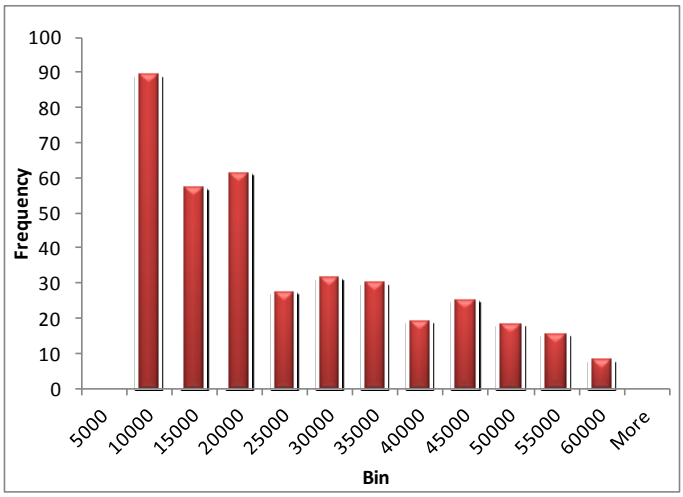


Figure 20. Office-Building distribution by square footage

Office applies to facility spaces used for general office, professional and administrative purposes. The two types of variables in the CBECS database are as follows:

- **Continuous:** refers to variables which have a real number or numeric value
- **Categorical:** refers to the variables which are discreet and have classes

Variables were chosen on the basis of their significance to electricity use. About 53 variables were selected from the entire list of variables from the CBECS database. Table 9 below gives a detailed list of the 53 variables. Out of these, 23 variables were shortlisted as the most important determinants of which, 7 were numeric.

Table 9.Description of Variables

S.no.	Variable	Description	Type	No. of Categories
1	YRCON	Year of Construction	Categorical	9
2	REGION	Region	Categorical	4
3	CENDIV	Census Division	Categorical	9
4	SQFT	Square feet	Numerical	
5	GLSSPC	Percent of Exterior Glass	Categorical	5
6	NFLOOR	Number of Floors	Categorical	9
7	BLDSHP	Building Shape	Categorical	11
8	WLCNS	Wall Constrcution Type	Categorical	9
9	RFCN	Roof Construction Type	Categorical	9
10	CLIMATE	Climate	Categorical	5
11	HDD	Heating Degree Days	Numerical	
12	CDD	Cooling Degree Days	Numerical	
13	HEATP	Percent Heated	Numerical	
14	COOLP	Percent Cooled	Numerical	
15	NWKR	Number of Workers	CAT/Numerical	12
16	WKHRS	Number of Working Hours	Categorical	7
17	PCTRMC	Number of PCs	CAT/Numerical	7
18	MAINCL	Main Cooling Equipment	Categorical	8
19	MAINHT	Main Heating Equipment	Categorical	7
20	VAV	Variable Air Unit	Categorical	2
21	ECN	Economizer	Categorical	2
22	LOHRPC	Percent lit when Open	Categorical	5
23	LNHRPC	Percent lit when Closed	Categorical	5
Additional Variables				
24	FURNAC8	Furnace	Categorical	2
25	BOILER8	Boiler	Categorical	2
26	PKGHT8	Packed Heating	Categorical	2
27	SLFCON8	Individual Heater	Categorical	2
28	HTPMPH8	Heat Pumps	Categorical	2
29	STHW8	District Steam	Categorical	2
30	OTHTEQ8	Other	Categorical	2
31	FURNP8	Furnace %	Numerical	
32	BOILP8	Boiler %	Numerical	
33	PKGHP8	Packed Heating %	Numerical	
34	SLFCNP8	Individual Heater %	Numerical	
35	HTPHP8	Heat Pumps %	Numerical	
36	STHWP8	District Steam %	Numerical	
37	OTHTP8	Other %	Numerical	
38	PKGCL8	Packaged Cooling	Categorical	2
39	RCAC8	Res Central A/C	Categorical	2
40	ACWNWL8	Indv A/C units	Categorical	2
41	HTPMPC8	Heat pumps	Categorical	2
42	CHWT8	Dist Chd wtr	Categorical	2
43	CHILLR8	Central chillers	Categorical	2
44	EVAPCL8	Evap coolers	Categorical	2
45	OTCLEQ8	Other	Categorical	2
46	PKGCL8	Packaged Cooling %	Numerical	
47	RCAC8	Res Central A/C %	Numerical	
48	ACWNWL8	Indv A/C units %	Numerical	
49	HTPMPC8	Heat pumps%	Numerical	
50	CHWT8	Dist Chd wtr %	Numerical	
51	CHILLR8	Central chillers %	Numerical	
52	EVAPCL8	Evap coolers %	Numerical	
53	OTCLEQ8	Other %	Numerical	

Table 10. EUI for entire dataset for Office buildings

Electricity used/Square footage	
Mean	15.66
Standard Error	0.81
Median	13.10
Standard Deviation	12.59
Minimum	0.63
Maximum	109.14
Count	242
Largest(1)	109.14
Smallest(1)	0.63
Confidence Level(95.0%)	1.59

Bin	*Frequency*
0	0
5	32
10	44
15	68
20	41
25	22
30	15
35	8
More	12

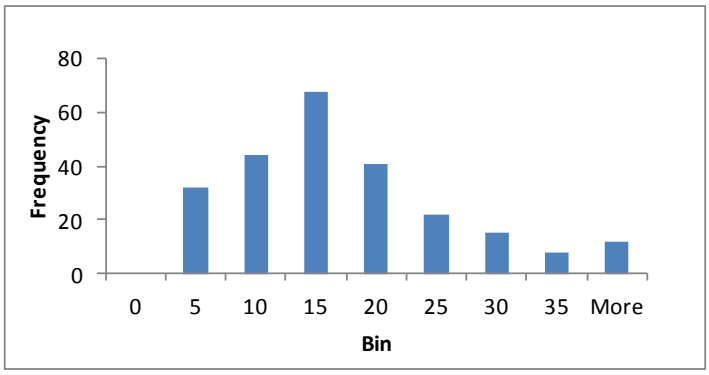


Figure 21. Graph to show EUI distribution for office buildings

The first step in the analysis is to plot the distribution of the response variable. It was found that the range of EUI for the data is from 0.63 to 110 kWh/Sq.ft. Looking at Table 10 and Figure 21 above, it is clear that the variability in the distributions is a result of the higher EUIs.

Normalizing the variables

Out of 23 variables, 7 were normalized to eliminate the effect of gross influences of certain variables (Table 11). The number of workers, number of Personal Computers (PCs) and EUI were normalized for 1000 sq.ft of floor area:

NWKR = Number of Workers/ (Sq.ft/1000)

PCSFT = Number of PCs/ (Sq.ft/1000)

Energy Use Intensity (EUI) = Electricity used/Sq.ft.

The Heating and cooling degree days were multiplied with percentage of area heated or cooled to normalize the conditioned area with climate. Table 11 shows the final list and descriptions of variables used in this study. After normalizing the variables, there are only 5 continuous or numeric variables out of 18 variables. The following are the calculations for the normalizations:

HDDPC = Heating Degree Days*Percent of floor space Heated

CDDPC = Cooling Degree Days*Percent of floor space Cooled

Table 11. Final list of selected variables

S.no.	Variable	Description	Type	No. of Categories
1	YRCON	Year of Construction	Categorical	9
2	CENDIV	Census Division	Categorical	9
3	GLSSPC	Percent of Exterior Glass	Categorical	5
4	NFLOOR	Number of Floors	Categorical	9
5	BLDSHP	Building Shape	Categorical	11
6	WLCNS	Wall Constrcution Type	Categorical	9
7	RFCN	Roof Construction Type	Categorical	9
8	HDDPC	Heating Degree Days*Percent Heated	Numerical	
9	CDDPC	Cooling Degree Days*Percent Cooled	Numerical	
10	NWKR	Number of Workers/(Sqft/1000)	Numerical	
11	WKHRS	Number of Working Hours	Categorical	7
12	PCSFT	Number of PCs/(Sqft/1000)	Numerical	
13	MAINCL	Main Cooling Equipment	Categorical	8
14	MAINHT	Main Heating Equipment	Categorical	7
15	VAV	Variable Air Unit	Categorical	2
16	ECN	Economizer	Categorical	2
17	LOHRPC	Percent lit when Open	Categorical	5
18	LNHRPC	Percent lit when Closed	Categorical	5
	EUI	Electricity used/Sqft	Numerical	

Outlier Removal

To eliminate the extreme EUI values the Inter-quartile range method mentioned in section 3.3 was adopted. Using that method twelve data points were detected and removed (Tables 12 and 13). Figure 22, is a graph depicting the observed EUI (response variable) plotted against the predicted EUI. Note that there are many high values of EUI. The circled data points were identified as the extreme outliers by the 'Statgraphics' tool (Figure 22). The values those were greater than the upper quartile which is 34.57kWh/sq.ft, were removed.

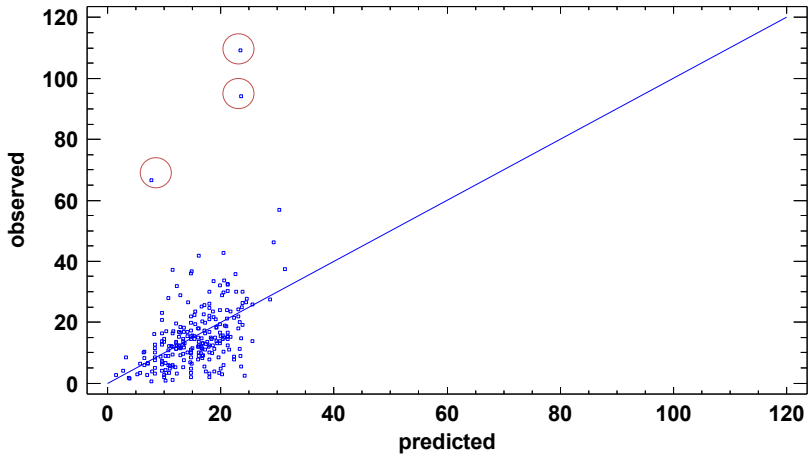


Figure 22. Outlier Detection based on EUI

Table 12. Outliers detected for Office buildings

Row	Y	Predicted Y	Residual	Studentized Residual
77	37.11	11.5601	25.5499	2.26
113	66.53	7.77853	58.7515	5.58
132	109.14	23.4846	85.6554	8.73
171	56.76	30.3235	26.4365	2.42
188	41.79	16.1265	25.6635	2.28
228	94.26	23.6054	70.6546	6.87

Table 13. The Inter-quartile range for Office buildings

Lower Quartile	**25th Percentile**	8.51
Upper Quartile	**75th Percentile**	18.94
Inter-Quartile Range	**IQR**	10.42
Lower quartile-1.5*IQR	**Bottom**	-7.1225298
Upper quartile+1.5*IQR	**Top**	34.5747087

Table 14 and Figure 23 below show that most of the data is covered within the first 4 bins and due to the presence of outliers the numbers of bins have increased.

Table 14. Descriptive Statistics of the range (Office buildings)

EUI with Outliers	
Mean	15.66
Standard Error	0.81
Median	13.10
Mode	13.02
Standard Deviation	12.59
Minimum	0.63
Maximum	109.14
Count	242.00

Bin	*Frequency*	*Cumulative %*
10	76	31.40%
20	109	76.45%
30	37	91.74%
40	13	97.11%
50	3	98.35%
60	1	98.76%
70	1	99.17%
80	0	99.17%
90	0	99.17%
100	1	99.59%
110	1	100.00%

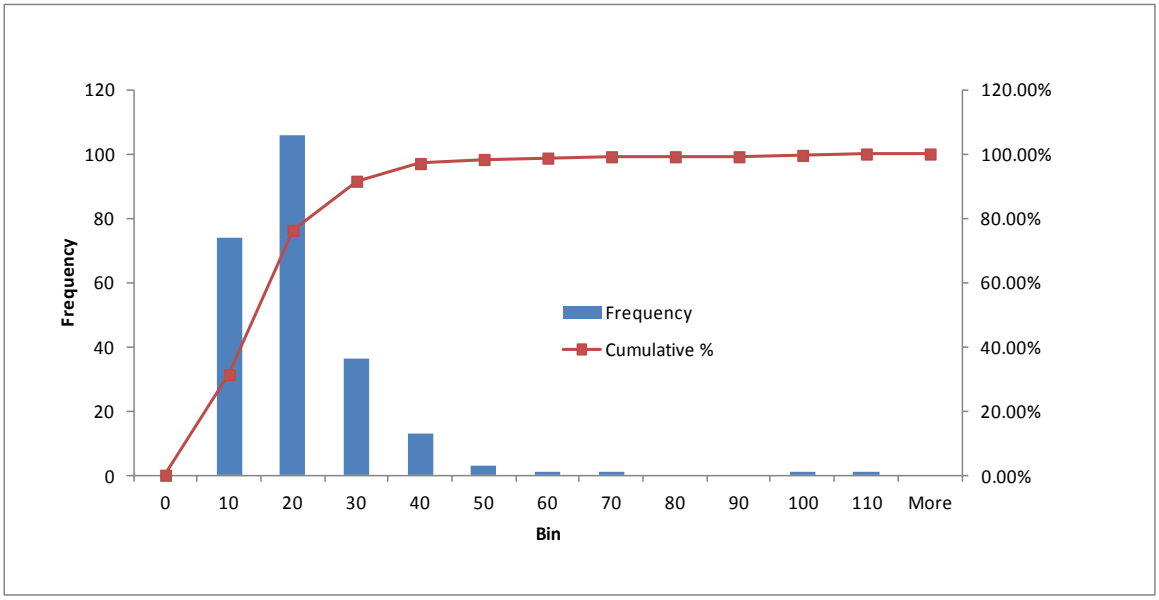


Figure 23. EUI distribution for the entire data (Office buildings)

EUI Data without Outliers

After the removal of outliers, the number of buildings eventually reduces from 242 to 230 buildings.

Table 15. Descriptive Statistics after removal of outliers for Office buildings

EUI without Outliers	
Mean	13.69
Standard Error	0.51
Median	12.67
Mode	13.02
Standard Deviation	7.69
Minimum	0.63
Maximum	33.82
Count	230.00

Bin	*Frequency*	*Cumulative %*
10	76	33.04%
20	109	80.43%
30	37	96.52%
40	8	100.00%

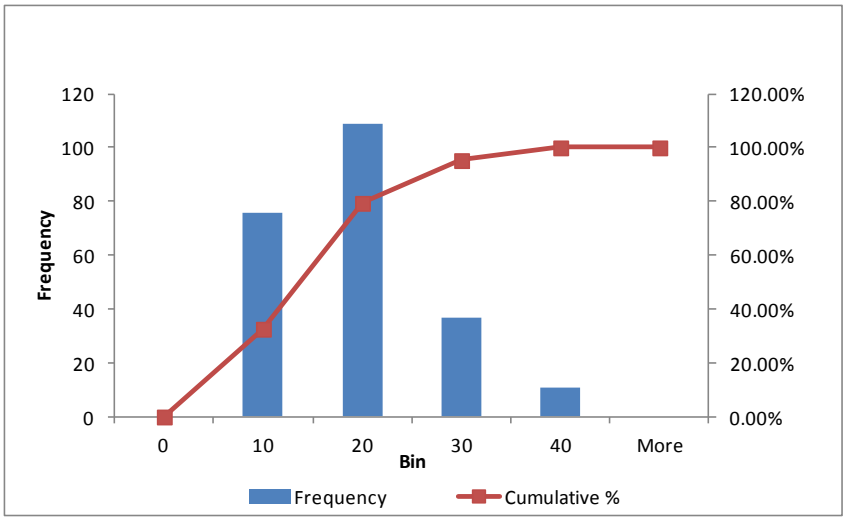


Figure 24. Graph for EUI distribution after the removal of outliers for Office buildings

There was a sizeable difference between the EUI medians and means. The mean value with outliers was about 15 kW/Sq.ft. (See Table 15). After removing the outliers, the

mean reduced to 13kW/Sq.ft. The problem of few excessively high EUIs significantly seems to increase the average EUI and there is very little influence on the median. (Sharp, 1996).

Coefficient of variation (CV) for this data: 80%.

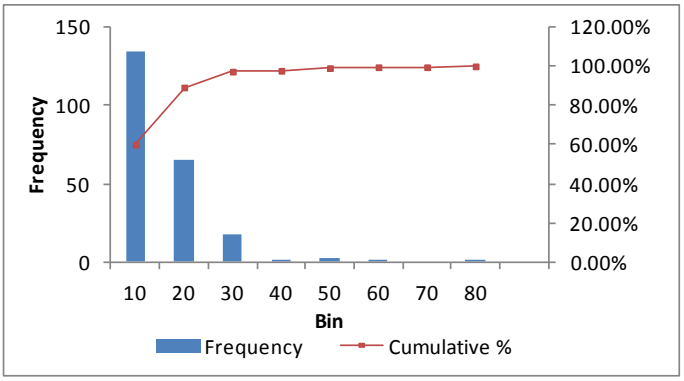
Coefficient of variation (CV) for this data (without outliers): 58%

4.1.2 School

School buildings were also extracted and filtered in the method mentioned in section 4.1.1. Table 16 gives the descriptive statistics for school buildings.

Table 16. Descriptive Statistics for School buildings without outlier removal

EUI distribution with outliers	
Mean	10.81
Median	8.30
Standard Deviation	9.38
Minimum	0.52
Maximum	79.38
Count	223



Bin	*Frequency*	*Cumulative %*
10	134	60.09%
20	65	89.24%
30	18	97.31%
40	1	97.76%
50	3	99.10%
60	1	99.55%
70	0	99.55%
80	1	100.00%

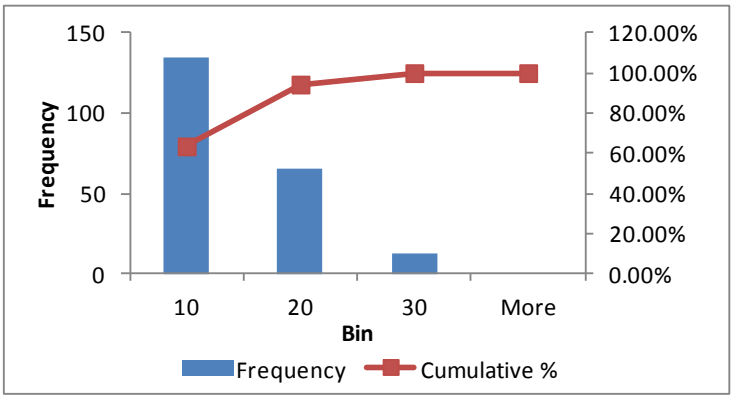
Figure 25. EUI distribution (with the presence of outliers) for School buildings

Removal of outliers

The outliers were removed as described in the previous section (Section 4.1.1). The mean reduces after the removal of outliers from 10.81kWh/Sq.ft. to 9.21 kWh/Sq.ft. Table 17 give the information of the data after the removal of outliers.

Table 17. Descriptive statistics for School buildings after outlier removal

EUI without Outliers	
Mean	9.21
Median	7.92
Standard Deviation	5.67
Minimum	0.52
Maximum	25.96
Count	211.00



Bin	*Frequency*	*Cumulative %*
10	134	63.51%
20	65	94.31%
30	12	100.00%
More	0	100.00%

Figure 26. Graph-EUI distribution for School buildings

4.2 Discretization methodology

Discretization is the process of transforming continuous variables into their discrete counterparts. This process is usually carried out as a first step towards making them suitable for numerical evaluation and implementation.

Table 18. Discretization for Office bldg. variables

S.no.	Variable	Description	Type	No. of Categories
1	YRCON	Year of Construction	Categorical	9
2	CENDIV	Census Division	Categorical	9
3	GLSSPC	Percent of Exterior Glass	Categorical	5
4	NFLOOR	Number of Floors	Categorical	9
5	BLDSHP	Building Shape	Categorical	11
6	WLCNS	Wall Constrcution Type	Categorical	9
7	RFCN	Roof Construction Type	Categorical	9
8	HDDPCcat	Heating Degree Days*Percent Heated	Categorical	4
9	CDDPCcat	Cooling Degree Days*Percent Cooled	Categorical	3
10	NWKR	Number of Workers/(Sqft/1000)	Categorical	3
11	WKHRS	Number of Working Hours	Categorical	7
12	PCSFTcat	Number of PCs/(Sqft/1000)	Categorical	4
13	MAINCL	Main Cooling Equipment	Categorical	8
14	MAINHT	Main Heating Equipment	Categorical	7
15	VAV	Variable Air Unit	Categorical	2
16	ECN	Economizer	Categorical	2
17	LOHRPC	Percent lit when Open	Categorical	5
18	LNHRPC	Percent lit when Closed	Categorical	5
	EUI	Electricity used/Sqft	Numerical	

As most of the variables were categorical, the new sets of normalized variables were converted into categorical variables. This was done to maintain uniformity in the data. For example, based on its distribution and range, HDDPH was converted to 4 classes (re-named as HDDPHcat) as follows:

Range		Class
0	2000	1
2001	4000	2
4001	6000	3
6001	10000	4

An analogous method was applied to CDDPH which was converted into 3 categories (re-named as CDDPHcat) as follows:

Range		Class
0	400	1
401	3000	2
3001	6000	3

Classification trees require the response to be a categorical variable. For this reason the response variable (i.e., EUI) was also discretized. See Appendix B for the other variables.

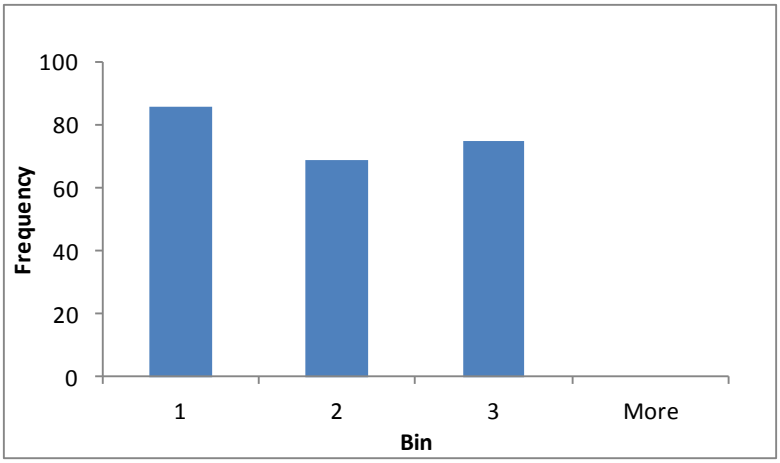
Categories for Office buildings data

The EUIs for office buildings were segregated into three and five categories (See Figure 27 &28) as given in Table 19 below. These classifications were based on the distribution of the data, and we wished to evaluate differences in our analysis results under each of these cases.

Table 19. Office buildings EUI range discretized in 3 and 5 categories

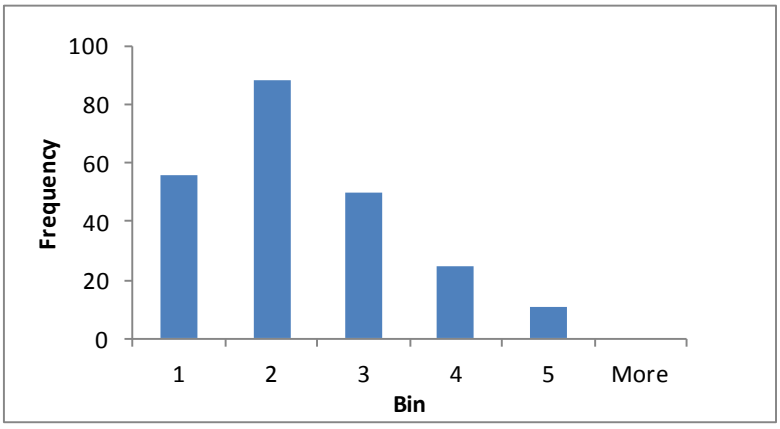
Range	Class	
0-10	1	Low
11 to 15	2	Med
16 to 35	3	High

Range	Class	
0-7	1	Low1
8 to 14	2	Low2
15 to 21	3	Med
22 to 28	4	High1
29 to 35	5	High2



Bin	*Frequency*
1	86
2	69
3	75
More	0

Figure 27. Distribution for three categories for Office buildings



Bin	*Frequency*
1	56
2	88
3	50
4	25
5	11
More	0

Figure 28. Distribution for five categories for Office buildings

The classes in each of the categorical regressor variables were reduced to 3 or 4 classes to minimize misclassification error rate, described earlier in Section 3.6. The methods used for category reduction are as follows:

(a). Create classes to have similar number of data points in each class

The categories are reduced in such a way that the numbers of data points in each category are more or less similar. For some variables such as equipment type, wall construction

and building shape, they were classified based on physical grouping. The following is an example of a variable classified according to the number of data points (See Figure 29 and Table 20&21):

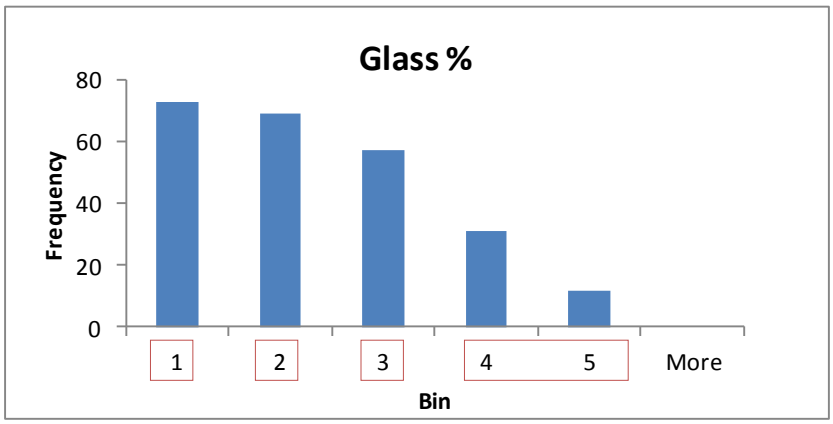


Figure 29. Distribution of Glass %

Table 20. Threshold of each class of the variable for office buildings

KEY	
1	10% or less
2	11% to 25%
3	26% to 50%
4	51% to 75%
5	76% to 100%

Table 21. Collapsing the classes for Office buildings

	Collapse to:	
1	1	10% or Less
2	2	11% to 25%
3	3	26% to 50%
4,5	4	51% to 100%

The following is an example of a variable classified according to type or physical grouping (Classification for all the variables is elaborated in Appendix B):

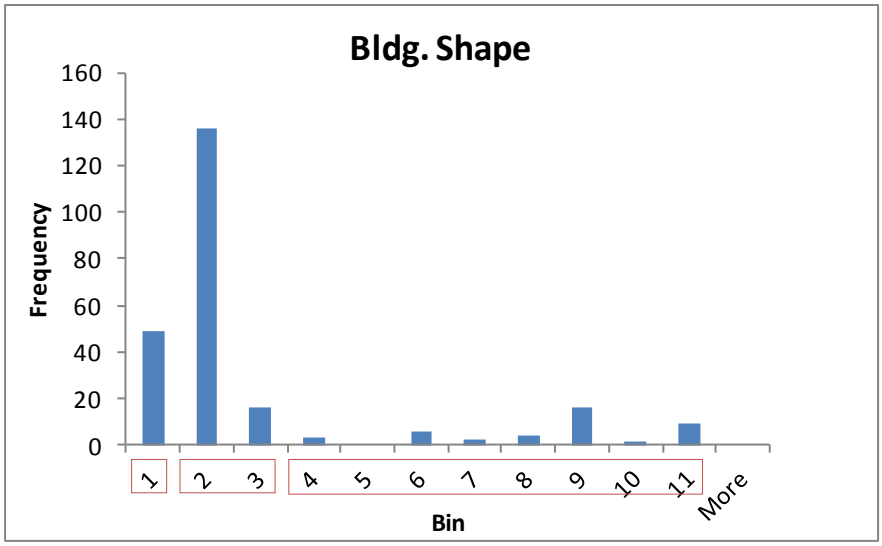


Figure 30. Distribution of classes for bldg shape

Table 22. Description of each class for Office buildings

KEY	
1	Square
2	Wide Rectangle
3	Narrow Rectangle
4	Rectangle/Square with courtyard
5	"H" Shaped
6	"U" Shaped
7	"E" Shaped
8	"T" Shaped
9	"L" Shaped
10	"+" Shaped
11	Other

Table 23. Collapsing the categories for Office buildings

	Collapse to:	
1	1	Square
2,3	2	Rectangular
4,5,6,7, 8,9	3	Other Shapes

In this example, shapes other than square and rectangle did not have as many data points and were grouped together and also that they could be classified as ‘Other shapes’.

(b). Combining similar categories (Schmueli et.al, 2011)

In this method, the distribution is divided in such a way that if the distribution were to be a normal distribution, the centre near the mean would be one class and the portion near the ends would be another class. An example of this method is shown below in Figure 31:

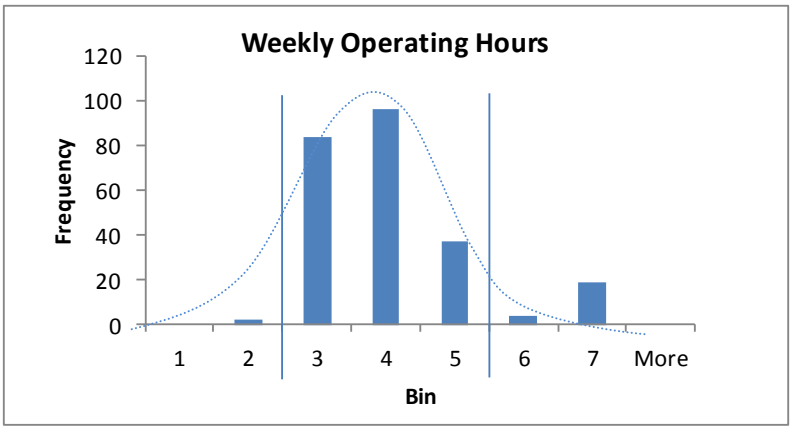


Figure 31. Combining similar classes for combined dataset

Table 24. Definition of classes for Working operating hours (in hours) for Office and School buildings

KEY	
1	Zero
2	1 to 39
3	40 to 48
4	49 to 60
5	61 to 84
6	85 to 167
7	Always open

Table 25. Collapsing the classes for Office and school buildings

	Collapse to:	
1,2	1	1 to 39
3,4,5	2	40 to 84
6,7	3	85 to 167 and more

4.3 Random Forest for office buildings

The random forest method was first applied to the medium office buildings dataset. The tool used for this method was MATLAB. The code was written to generate 500 trees. The average values of all the 500 trees for MSE and CV or ensemble error were provided in the end as a result. For the regression version of random forest the results are best understood with the performance metrics. The metrics for the random forest (regression version) for office buildings were-

- Mean Squared Error (MSE): 40.04
- Coefficient of Variation: 0.46 (or 46%).

The following were the results for the classification version of random forest as a confusion matrix below:

Table 26. Confusion Matrix for 3 categories

Class	Med	High	Low	Row Total	Model Error
Med	27	22	20	**69**	61%
High	18	43	14	**75**	43%
Low	9	11	66	**86**	23%
Col Total	**54**	**76**	**100**	**230**	
Use Error	50%	43%	34%	**40.87%**	
				Ensemble error	

Here, the 'Medium' category was classified correctly only 27 out of 69 numbers. The numbers in the diagonal boxes give the correct predictions. 'Use error' specifies thee misclassification of other classes as a given class. 'Model error' gives the total error of misclassification of a particular class as another class. To compare the performance of the

two versions, the CV values for the ensemble errors were calculated. The CV for EUI classification of 3 categories was found to be 0.59 (59%).

Table 27. Confusion Matrix for 5 categories

Class	High2	High1	Med	Low1	Low2	Row Total	Model Error
High2	14	27	2	7	0	**50**	72%
High1	10	64	14	0	0	**88**	27%
Med	6	24	24	2	0	**56**	57%
Low1	13	8	2	2	0	**25**	92%
Low2	2	8	0	1	0	**11**	100%
Col Total	**45**	**131**	**42**	**12**	**0**	**230**	
Use Error	69%	51%	43%	83%	0%	**54.78%**	
						Ensemble error	

The CV for random forest (classification version) with response variable having 5 categories is: 0.56(56%). The improvement is only 3%, but, this method was mainly used to find the important variables or building parameters that would have a dominant influence on the EUI in the dataset.

4.4 Combining Office and School data

The above model did not perform very well with the office data. The numbers of buildings for office buildings were 230, which are too few data points. A possible approach to improve the performance was to increase the size of the sample. We decided to combine the office and school buildings data to increase the number of data points. The total number of buildings now was 441 without outliers. This new dataset was used for the random forest technique (regression and classification version).

Table 28. Description statistics for combined data

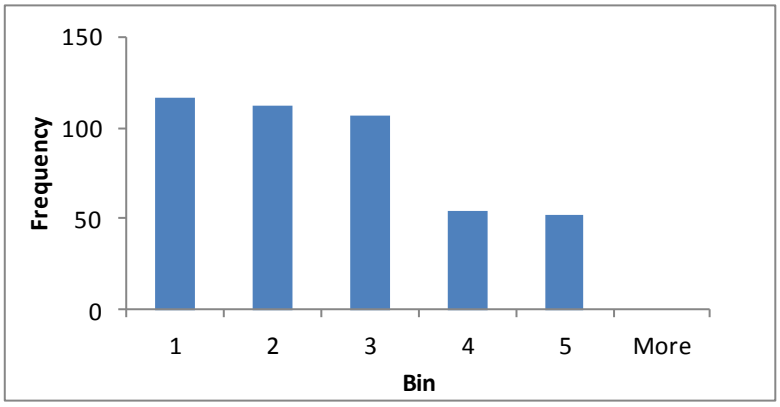
EUI for combined data	
Mean	11.55
Median	10.44
Standard Deviation	7.15
Minimum	0.52
Maximum	33.82
Sum	5093.03
Count	441
Largest(1)	33.82
Smallest(1)	0.52

The data was discretized and the number of categories for each variable and EUI were reduced using methods mentioned in section 4.2.

EUI segregated into 5 categories:

Table 29. Description of 5 categories for combined data

EUI Categories			
Name	**Range**		**Class**
Lowest	0	5	1
Low	6	10	2
Medium	11	15	3
High	16	20	4
Highest	21	35	5



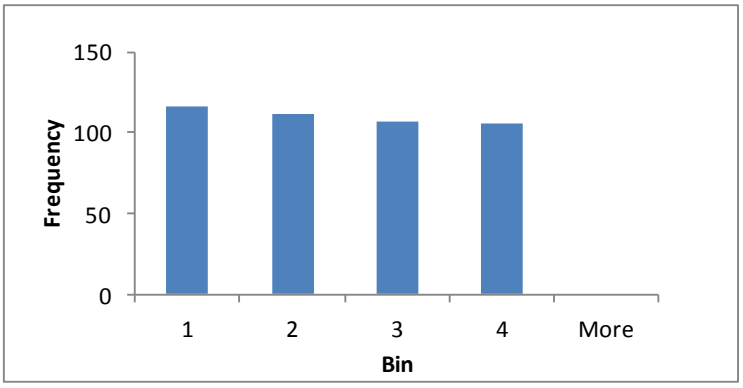
Bin	*Frequency*
1	116
2	112
3	107
4	54
5	52
More	0

Figure 32. Distribution of EUI with 5 categories for combined data

EUI segregated into 4 categories

Table 30. Description of 4 categories for combined data

EUI Categories			
Name	**Range**		**Class**
Lowest	0	5	1
Low	6	10	2
Medium	11	15	3
High	16	35	4



Bin	*Frequency*
1	116
2	112
3	107
4	106
More	0

Figure 33. Distribution of EUI with 4 categories for combined data

Both these EUI categories were used to for the classification version of random forest. These categories were segregated according to the method mentioned in section 4.2 (b).

4.5 Random Forest for combined building data

The procedure carried out with office buildings was now executed using the combined dataset (see section 4.3). The result for the regression version is best. Summarized by the CV value of 0.49 (49%).

The result, for the classification version of random forest: result in a CV value of 0.52(52%) for 5 categories of EUI and a CV value of 0.56 (56%).

Table 31. Confusion matrix for the other method of classification into 5 categories

Class	Lowest	Low	Med	High	Highest	Row Total	Model Error
Lowest	64	33	15	1	3	**116**	45%
Low	39	39	28	2	4	**112**	65%
Med	20	26	46	7	8	**107**	57%
High	5	14	26	9	0	**54**	83%
Highest	3	8	19	3	19	**52**	63%
Col Total	**131**	**120**	**134**	**22**	**34**	**441**	
Use Error	51%	68%	66%	59%	0%	**59.86%**	
						Ensemble error	

Table 32. Confusion matrix for 4 categories

Class	Lowest	Low	Med	High		Row Total	Model Error
Lowest	67	30	13	6		**116**	42%
Low	32	39	26	15		**112**	65%
Med	16	24	39	28		**107**	64%
High	7	15	28	56		**106**	47%
Col Total	**122**	**108**	**106**	**105**		**441**	
Use Error	45%	64%	63%	47%		**54.42%**	
						Ensemble error	

The models did not result in much better predictive ability but, the random forest model (classification version) of the combined dataset, performed better than the classification version of the office dataset by itself.

4.6 Identifying important Variables

The variable importance was obtained from the random forest method. The higher the value of the z-scores, higher will be the variable ranking. The variable with highest z-score or the most dominant variable was considered the base (100%) and the other variables were divided by the base to provide the percentage importance of the variables.

For Office Buildings

Regression Version

Table 33. Variable Importance

	Name	Z scores	Rank
1	'CENDIV8	0.128627	1
2	'PCSFTcat'	0.095997	0.75
3	'MAINCL'	0.090174	0.70
4	'NWSFTca	0.081551	0.63
5	'GLSSPC'	0.064129	0.50
6	'VAV8 '	0.059348	0.46
7	'LOHRPC'	0.054786	0.43
8	'MAINHT'	0.051031	0.40
9	'YRCONC'	0.050639	0.39
10	'HDDPHca	0.050293	0.39
11	'NFLOOR'	0.045411	0.35
12	'RFCNS'	0.043066	0.33
13	'CDDPCcat	0.03999	0.31
14	'LNHRPC'	0.039777	0.31
15	'ECN8 '	0.036268	0.28
16	'BLDSHP'	0.035753	0.28
17	'WKHRSC'	0.031145	0.24
18	'WLCNS'	0.030922	0.24

KEY	
'YRCONC'	Year of Construction
'CENDIV8 '	Census Division
'NWSFTcat'	No of Workers/1000 Sqft
'WKHRSC'	Working Hours/week
'NFLOOR'	No. of Floors
'GLSSPC'	Exterior Glass Percentage
'BLDSHP'	Building Shape
'WLCNS'	Wall Construction
'RFCNS'	Roof Construction
'HDDPHcat'	Heating Degree Day*Percentage Heated
'CDDPCcat'	Cooling Degree Day*Percentage Cooled
'MAINHT'	Main Heating Equipment
'MAINCL'	Main Cooling Equipment
'VAV8 '	Variable Air Volume
'ECN8 '	Economizer
'PCSFTcat'	No. of PCs/1000 Sq Ft.
'LOHRPC'	% Lit when Building Open
'LNHRPC'	% Lit when Building Closed

The most important variables in this method are the census division, number of PCs/1000 sq.ft. cooling equipment, number of workers/1000sq.ft and glass %. Note that the census division has highest z-score and is assumed as the base. The ranks for other variables are

calculated by normalizing them with the base. See Table 33 for the z-scores and ranks. Figure 34 shows the variable ranking.

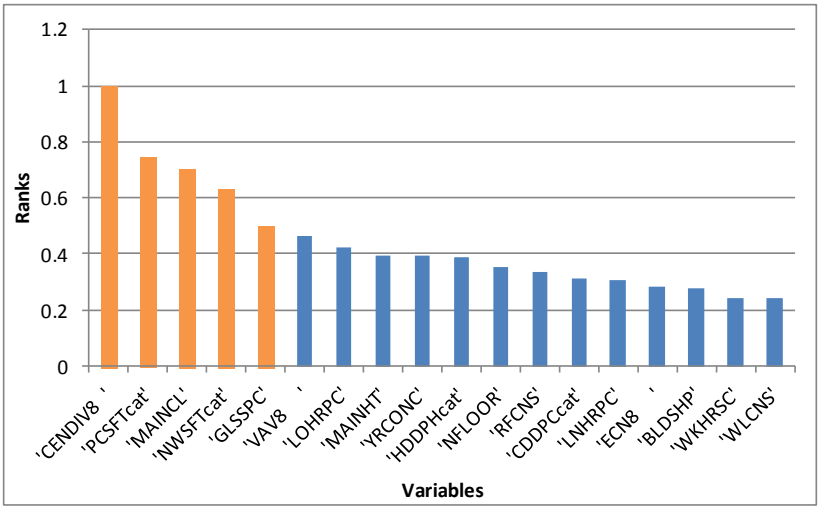


Figure 34. Graph depicting the important variables of Office buildings (Regression version)

Classification Version

Random forest classification version for EUI with 5 categories performed better than the data which had 3 categories in EUI. So, variable ranking for EUI with 5 categories is given below in Table 34 and Figure 35. Table 34 shows the z-scores and variable ranking for Office buildings obtained from random forest classification version. Figure 35 depicts this variable ranking. The other variables were ranked in a method similar to the one used for the regression version for Office buildings. The base variable here also was census division. Number of floors and the year of construction also are some of the important variables here.

Table 34. Ranking of variables

Name	Zs	Rank
'CENDIV8	0.000898	1
'GLSSPC'	0.000654	0.73
'MAINHT'	0.000544	0.61
'NFLOOR'	0.000524	0.58
'YRCONC'	0.000519	0.58
'PCSFTcat'	0.000518	0.58
'HDDPHca	0.000497	0.55
'MAINCL'	0.000491	0.55
'RFCNS'	0.000463	0.52
'NWSFTca	0.000439	0.49
'LNHRPC'	0.000416	0.46
'WKHRSC'	0.000416	0.46
'BLDSHP'	0.000406	0.45
'WLCNS'	0.000368	0.41
'ECN8 '	0.000365	0.41
'VAV8 '	0.000318	0.35
'CDDPCcat	0.000313	0.35
'LOHRPC'	0.000308	0.34

KEY	
'YRCONC'	Year of Construction
'CENDIV8 '	Census Division
'NWSFTcat'	No of Workers/1000 Sqft
'WKHRSC'	Working Hours/week
'NFLOOR'	No. of Floors
'GLSSPC'	Exterior Glass Percentage
'BLDSHP'	Building Shape
'WLCNS'	Wall Construction
'RFCNS'	Roof Construction
'HDDPHcat'	Heating Degree Day*Percentage Heated
'CDDPCcat'	Cooling Degree Day*Percentage Cooled
'MAINHT'	Main Heating Equipment
'MAINCL'	Main Cooling Equipment
'VAV8 '	Variable Air Volume
'ECN8 '	Economizer
'PCSFTcat'	No. of PCs/1000 Sq Ft.
'LOHRPC'	% Lit when Building Open
'LNHRPC'	% Lit when Building Closed

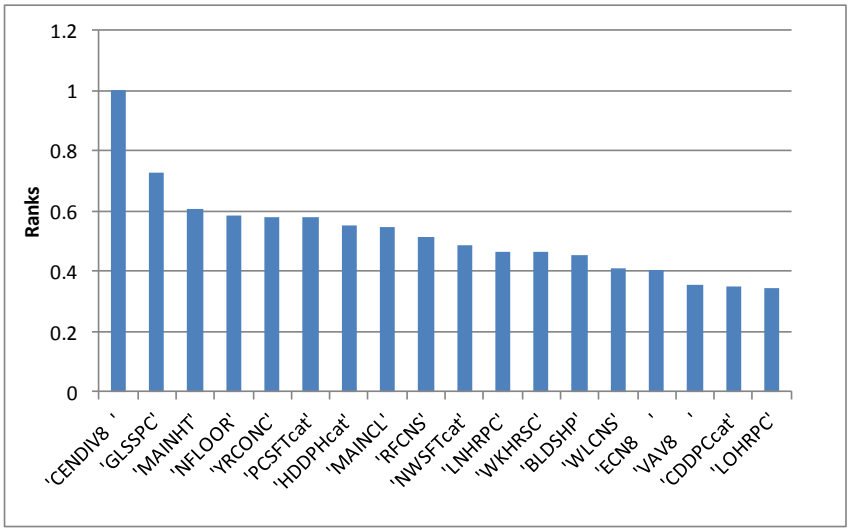


Figure 35. Graph depicting the important variables of Office buildings (Classification version)

For Combined data

Regression Version

Table 35. Variable importance for regression ensemble

Name	Z-scores	Rank
'CENDIV8 '	0.043611	1
'MAINCL9'	0.042849	0.98
'PCSFTcat'	0.034996	0.80
'NWSFTcat'	0.032973	0.76
'ECN8 '	0.030627	0.70
'CDDPCcat'	0.030492	0.70
'MAINHT9'	0.026971	0.62
'VAV8 '	0.024259	0.56
'LNHRPC9'	0.021413	0.49
'RFCNS9'	0.018744	0.43
'BLDSHP9'	0.018125	0.42
'LOHRPC9'	0.017495	0.40
'YRCONcat'	0.015821	0.36
'HDDPHcat'	0.015257	0.35
'GLSSPC9'	0.015086	0.35
'WLCNS9'	0.01492	0.34
'WKHRSCat'	0.013891	0.32
'NFLOOR9'	0.012272	0.28

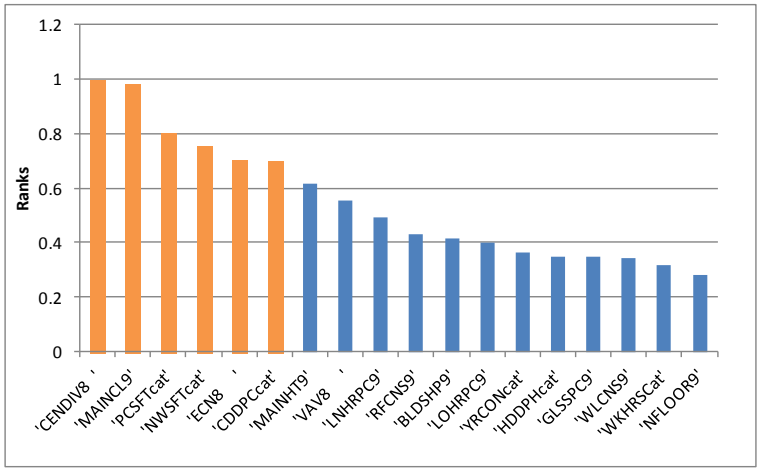


Figure 36.Important variables in combined dataset (Regression version)

Table 35 and Figure 36 give the variable ranking for random forest regression version. Note that the census division has the highest z-score and so was considered as the base variable and the other variable ranking was found using the base. Here the other important variables are main cooling equipment, number of PCs per 1000sq.ft, number of workers, economizer and cooling degree days.

Classification Version:

Table 36 and Figure 37; give the variable ranking for 5 categories, since that model performed better than the one with 3 categories alone. The ranking for other variables was found in the method mentioned for regression version.

Table 36. Variable importance for classification ensemble

Name	Z-scores	Rank
'CENDIV8 '	0.000396	1
'MAINCL9'	0.000245	0.62
'MAINHT9'	0.000245	0.62
'PCSFTcat'	0.000234	0.59
'RFCNS9'	0.000231	0.58
'BLDSHP9'	0.000220	0.55
'LNHRPC9'	0.000214	0.54
'NWSFTcat'	0.000204	0.52
'CDDPCcat'	0.000202	0.51
'YRCONcat'	0.000175	0.44
'GLSSPC9'	0.000174	0.44
'HDDPHcat'	0.000174	0.44
'WLCNS9'	0.000172	0.43
'ECN8 '	0.000171	0.43
'VAV8 '	0.000130	0.33
'WKHRSCat'	0.000103	0.26
'LOHRPC9'	0.000096	0.24
'NFLOOR9'	0.000094	0.24

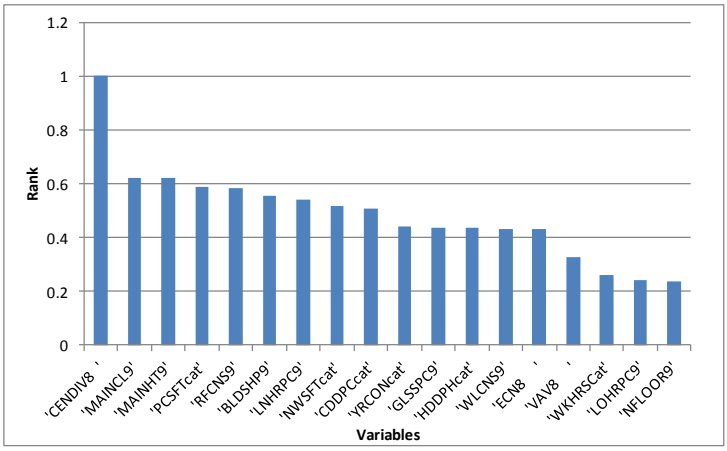


Figure 37.Important variables for combined dataset (Classification version)

4.7 Comparative analysis with prior work

The last step in analysis was to compare the result of the above model with an existing benchmarking model which used the same database. The OLS method used by ENERGY STAR Portfolio Manager uses the CBECS database (See Section 2.3.1 for the method used in Portfolio Manager). This method uses only continuous variables from the database. This resulted in the tool utilizing only 6 variables.

For this step, same set of continuous variables from the dataset (i.e. Office and Office & School) chosen for this research, were used to generate an OLS model. The variables were: floor area (Sq.ft.), number of workers, working hours, heating/cooling degree days, percentage of area heated /cooled, number of personal computers and EUI as the response

variable. The procedure for normalizing the continuous variables was adopted from the Portfolio Manager and applied to the variables used from the dataset of this study (See Table 37):

Table 37. Normalizing the continuous variables (ENERGY STAR, Portfolio Manager)

LN(Sft)	Natural Log of floor area (square footage)
LN(Nwk Den)	Natural Log of number of workers per 1000 Sq.ft.
LN(WKHR)	Natural Log of number of working hours
HDD*(HP/100)	Heating Degree Days*Percentage of area heated
CDD*(CP/100)	Cooling Degree Days*Percentage of area cooled
PC/(Sft/1000)	Number of personal computers per 1000 Sq.ft.
Elec/Sqft	Electricity used per Square Feet

A stepwise regression was carried out in a statistics tool called SPSS. Using the normalized variables mentioned earlier. The performance of the model was then compared to that of the model of the Portfolio Manager. The results for each dataset are discussed below:

Linear Regression for Office buildings

Table 38 gives a summary of results of the regression analysis. R^2 for this model was found to be 0.31 (31%). This indicates that the model can explain only 31% of the variance in EUI for Office buildings. The model developed by Portfolio Manager had an R^2 of 0.33(33%). These results suggest that the linear regression model developed using the data for this study is similar to that used by Portfolio manager.

Table 38. Model Summary for Office Building data

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.453[a]	.205	.202	6.87048
2	.497[b]	.247	.240	6.70283
3	.534[c]	.285	.276	6.54490
4	.546[d]	.298	.286	6.49976
5	.559[e]	.313	.297	6.44655

Inter-comparison of the models should be based on a common performance metric. So far the coefficient of variation (CV) has been used to evaluate the modeling accuracy or performance of the model. In this case, standard error with the mean of the model being 13KWh/sq.ft., CV of this model is 0.49 (49%). The coefficients for this model have been given in Table 39.

Table 39. Coefficients for the model

Model		Unstandardized Coefficients		t	Sig.
		B	Std. Error		
1	(Constant)	11.463	.538	21.305	.000
	LNNwkden	4.279	.557	7.676	.000
2	(Constant)	9.780	.708	13.812	.000
	LNNwkden	4.186	.544	7.689	.000
	CDDCP100	.001	.000	3.542	.000
3	(Constant)	7.798	.896	8.702	.000
	LNNwkden	2.654	.691	3.843	.000
	CDDCP100	.001	.000	3.732	.000
	PCSft1000	1.313	.378	3.477	.001
4	(Constant)	-14.440	10.953	-1.318	.189
	LNNwkden	2.450	.693	3.535	.000
	CDDCP100	.001	.000	3.707	.000
	PCSft1000	1.445	.381	3.797	.000
	LNSqft	2.150	1.055	2.037	.043
5	(Constant)	-19.951	11.155	-1.789	.075
	LNNwkden	2.407	.688	3.500	.001
	CDDCP100	.002	.000	4.284	.000
	PCSft1000	1.518	.379	4.005	.000
	LNSqft	2.404	1.053	2.282	.023
	HDDHP100	.000	.000	2.175	.031

a. Dependent Variable: ElecSqft

Linear Regression for Combined data

The analysis described above was repeated using the dataset for school and office buildings. Four different models were evaluated (Table 40 and 41) and model 4 is

selected as the best one. The R^2 of the model is 0.37 and the CV is 0.48. The results are quite close to the R^2 value of the model used by Portfolio Manager which is 0.33(33%).

Table 40. Model Summary for combined data

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.491[a]	.241	.240	6.23574
2	.568[b]	.323	.320	5.89664
3	.596[c]	.355	.351	5.76275
4	.612[d]	.375	.369	5.68149

Table 41. Coefficients for the model for combined data

Model		Unstandardized Coefficients		Standardized Coefficients	t
		B	Std. Error	Beta	
1	(Constant)	10.680	.306		34.911
	LNNwkden	4.365	.369	.491	11.818
2	(Constant)	8.612	.406		21.237
	LNNwkden	4.254	.350	.479	12.167
	CDDCP100	.002	.000	.286	7.276
3	(Constant)	6.810	.554		12.283
	LNNwkden	3.317	.397	.373	8.363
	CDDCP100	.002	.000	.293	7.602
	PCSft1000	.989	.213	.207	4.646
4	(Constant)	-5.048	3.263		-1.547
	LNNwkden	3.078	.396	.346	7.766
	CDDCP100	.002	.000	.300	7.901
	PCSft1000	1.055	.211	.221	5.009
	LNWkHrs	2.925	.794	.142	3.686

a. Dependent Variable: ElecSqft

Linear Regression for Office buildings using Sharp's model

The model proposed by Sharp (See section 2.4 for the discussion about this model) was then applied to the medium office buildings selected for this study. This was another step for cross-verification of the data. The variables used in this model were identified in the

dataset of medium office buildings. The variables were: logarithm of number of workers per square feet, number of personal computers, occupancy type, working hours, economizer, and chiller. Occupancy type was a variable which was available in CBECS 1999.Occupancy type informs whether the building is owner occupied or not. This variable was removed from the CBECS 2003 so this variable was eliminated for this procedure. The buildings with missing values were deleted. There were 242 buildings. After the removal of outliers and missing vales, there were 225 buildings. A stepwise regression was run and the following results were acquired:

Table 42.Model summary for medium Office buildings (based on Sharp's model)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.444[a]	.197	.194	6.87
2	.533[b]	.284	.277	6.50

Table 43. Coefficients for medium Office buildings (based on Sharp's model)

Model		Unstandardized Coefficients		Standardized Coefficients	t
		B	Std. Error	Beta	
1	(Constant)	41.423	3.746		11.057
	LogNwkersft	9.961	1.345	.444	7.406
2	(Constant)	51.641	4.058		12.724
	LogNwkersft	9.516	1.276	.424	7.457
	CHILLR8	-6.206	1.198	-.295	-5.179

Two different models were assessed (Table 42 &43). Model 2 performed better than the other model. The R^2 of this model was found to be 0.27 and the CV was 0.46.

4.8 Model Accuracy

This section compares the different models developed in this study. The results are given in the following sections.

4.8.1. For Office Buildings:

Table 44 gives the coefficient of variation for the different models evaluated: mean model (which is the CV of the data without any model), linear regression model and random forest (regression and classification version) model for Office buildings alone.

Table 44. Comparison of Models

Method	Coefficient of Variation (CV)
Mean Model	57%
Linear Regression Model	49%
RF Regression version	46%
RF Classification version	3 categories: 59%
RF Classification version	5 categories: 55%

It is found that the model that performs the best is the random forest regression version. For visual interpretability, with the important variables generated by this method was selected and used to build a single regression tree. Figure 38 is the single tree built using the top five important variables namely: Number of personal computers, location, main cooling equipment, number of workers, exterior glass %. The order of importance is not necessarily the same as the random forest averages the z-scores for all the 500 trees.

4.8.2. For Office and School Buildings (Combined Data)

Using the combined data, it was found that random forest regression version and linear regression models perform better than the classification version of random forest.

Table 45. Comparison of models

Method	Coefficient of Variation (CV)
Mean Model	62%
Linear Regression Model	48%
RF Regression version	49%
RF Classification version	5 categories: 52%
RF Classification version	4 categories: 56%

For conceptual interpretability a similar single regression tree was generated using the top six important variables from Table 35 of section 4.6 (Figure 39).

The root node is the variable that is the most influential determinant followed by the variables at the leaf nodes. The terminal nodes give the EUI for that branch. For example, in Figure 38, a building that falls in category 1 of number of PCs per 1000 sq. ft, and lies in the location specified by categories 1, 2 and 9of census division, will have a low EUI of 7. This can be further understood by noting that Category 1 of number of PCs has upto 20 PCs which are very few and locations 1, 2 and 9 are New England, middle Atlantic and Pacific respectively which do not need excessive mechanical cooling, have low EUI. We note from Figure 38 that the first branching is based on the number of PCs, then by Census division and so on. The tree has 4 layers with exterior glass% being the last variable at which splitting is done. If on the other hand, a tree with 6 of the most important variables for office and school buildings combined is selected (Figure 39), the splitting changes quite drastically with “economizer” being the primary splitting variable, and Census Division being the last one. The order of variable importance for a single tree and random forest are not necessarily the same because random forest result is an average of 500 trees. The single tree is a representation of one of those 500 trees.

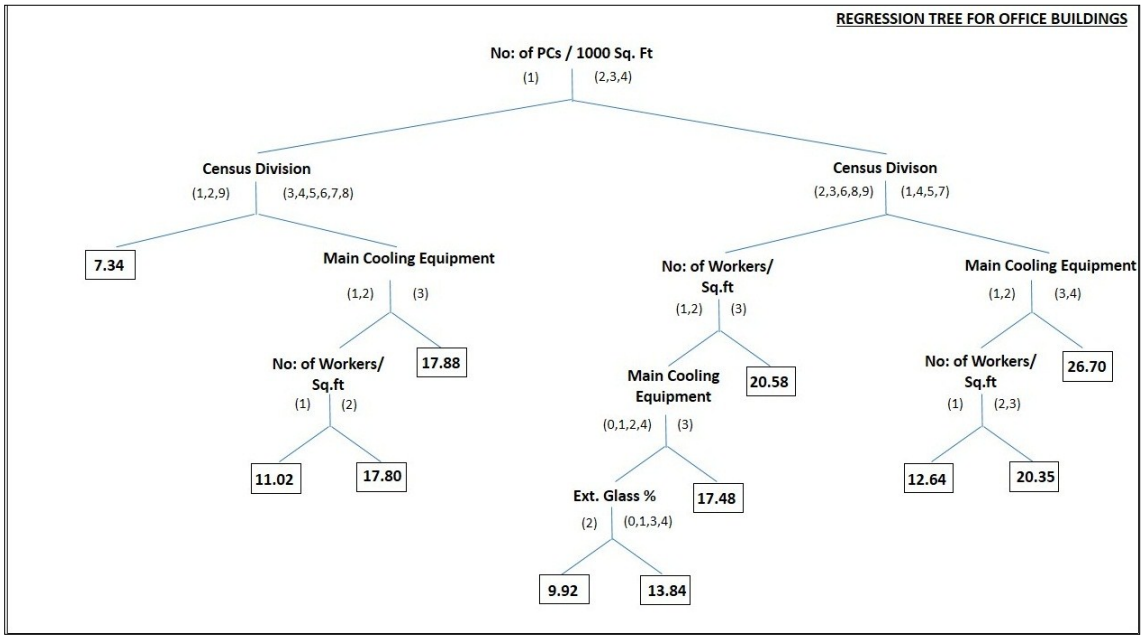


Figure 38. Single regression tree for office buildings using 5 important variables

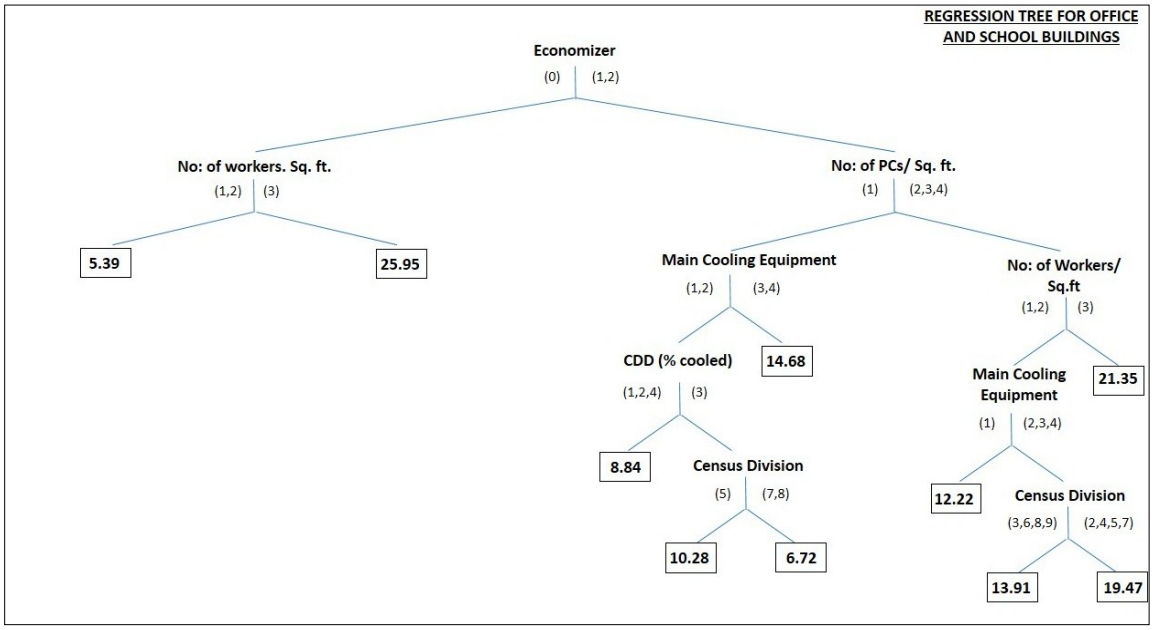


Figure 39. Single regression tree using 6 most important variables for office and School buildings combined

The single tree is not a robust model and is used only for conceptual interpretability. On the other hand, despite being an unintelligible model, random forest is a robust model and hence can be used for determining variable importance and the single tree is used just for a conceptual understanding of the strong determinants.

CHAPTER 5

SUMMARY AND FUTURE WORK

5.1 Summary

Benchmarking a building for its energy use gives the building owner and the facility manager a fair understanding about the energy use efficiency of the building compared to its peer group and its potential for energy savings, thereby decreasing their continuing costs of operation in terms of utility bills.

This study proposed a new benchmarking approach based on decision trees, where a relationship between the energy use intensities (EUI) and building parameters was established for two building types (office and school). The data for office and school buildings was extracted and filtered from the CBECS database. The Random forest technique was used to find the most influential parameters on building energy use intensities.

Out of the two methods of random forest used, the regression version performed better than the classification version. The top 5 or 6 variables were chosen from this method and a single regression tree was built for conceptual interpretability. The modeling approach proposed and evaluated in this study is only slightly better than the current benchmarking models. For example, the CV values for the office and school buildings combined, improved by 3% (from 49% to 46%) only. However, the general approach is much more

systematic and evaluates the effect of the numerous categorical variables contained in the CBECS database.

From the simplified models (single regression tree) we identified that location, system type (and other system related details), percentage of exterior glazing or information about the envelope are important determinants of energy use intensities. These are not important criteria used in many benchmarking tools. The reason for this discrepancy is worthy of future investigation.

5.2 Limitations

After studying and calculating indices from the CBECS survey database it is found that there are many unusual characteristics for many office and school buildings in the survey. Some of these include extreme EUI values, excessive or minimal square footage per worker, and building where the reported floor areas were greater than the calculated floor areas. Building characteristics that are known to be important determinants of energy use such as lighting wattage, information regarding the system type and building heat transfer coefficient are unavailable in the database.

This research uses the CBECS 2003 database which needs to be improved and updated. This survey was conducted many years ago. All benchmarking tools presently use the CBECS 2003 database. The development of CBECS 2012 is in progress and new variables and details are being added.

5.3 Future Research

The regression tree methodology is marginally better than the linear regression models used in current benchmarking tools. However refinements to the general approach suggested in this thesis is likely to improve in terms of prediction accuracy.

This study could be extended to using other databases for other building typologies with different weather conditions to better analyze the performance and the effectiveness of the model. The model could then be compared with the other methods used by the benchmarking tools in the market.

A further study of tree-ed regression technique with the important variables (obtained in the random forest method) could be conducted to achieve an intuitive model. Also, a synthetic study to evaluate whether the tree-ed regression technique would be a better prediction model than linear regression model would be an interesting avenue for research.

Further, evaluating this method using the newly released version of the CBECS 2012 would give further understanding of how newly added building variables are likely to influence the EUI.

REFERENCES

About the commercial buildings energy consumption survey. (2013). Retrieved 7/10, 2013, from http://www.eia.gov/consumption/commercial/about.cfm

Agnihotri, S. P. (2011). *Comparative analysis of benchmarking and audit tools*. Tempe, AZ: MSBE Thesis, The Design School, Arizona State University.

Architecture 2030: A historic opportunity. (2010). Retrieved 7/16, 2013, from http://architecture2030.org/the_solution/buildings_solution_how

ASHRAE Building Energy Labeling Program Implementation Committee. (2009). *ASHRAE Building Energy Labeling Program*. Atlanta, GA: ASHRAE Building Energy Labeling Program.

ASHRAE Building Energy Quotient Program. (2009). Value of ASHRAE's Building EQ Program. Retrieved 7/11, 2013, from http://www.buildingeq.com/index.php/about1abel

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32.

Breiman, L. (2003). *RF/tools: A class of two-eyed algorithms*. SIAM Workshop, Statistics Department, UC Berkeley.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Monterey, CA: Wadsworth & Brooks.

Buildings Technology Center - ORNL. (1996). Estimate Savings Potential. Retrieved 6/10, 2013, from http://eber.ed.ornl.gov/benchmark/est.htm

Kong, C. S., Villars, P., Iwata, S., & Rajan, K. (2012). Mapping the 'materials gene' for binary intermetallic compounds—a visualization schema for crystallographic databases. *Computational Science & Discovery*, 5(1), 015004.

CBECS. (2003). 2003 CBECS Survey Data. Retrieved 12/3, 2012, from http://www.eia.gov/consumption/commercial/data/2003/

Chung, W. (2011). Review of building energy-use performance benchmarking methodologies. *Applied Energy, 88*(5), 1470-1479.

Chung, W., & Hui, Y. V. (2009). A study of energy efficiency of private office buildings in hong kong. *Energy and Buildings, 41*(6), 696-701.

Chung, W., Hui, Y. V., & Lam, Y. M. (2006). Benchmarking the energy efficiency of commercial buildings. *Applied Energy, 83*(1), 1-14.

Classification and regression trees. (2009). Retrieved 6/15, 2013, from http://www.stat.cmu.edu/~cshalizi/350/lectures/22/lecture-22.pdf

Clean and affordable energy act (CAEA) of 2008. (2008). Retrieved 7/3, 2013, from http://green.dc.gov/sites/default/files/dc/sites/ddoe/publication/attachments/CAEA_of_2008_B17-0492.pdf

Clifton, C. (2010). Encyclopædia Britannica: Definition of Data Mining. Retrieved 7/4, 2013, from http://www.britannica.com/EBchecked/topic/1056150/data-mining

Coefficient of variation. (2013). Retrieved 7/2, 2013, from http://en.wikipedia.org/wiki/Coefficient_of_variation

Cross-validation (statistics). (2013). Retrieved 7/1, 2013, from http://en.wikipedia.org/wiki/Cross-validation_(statistics)

DDOE. (2013). Energy benchmarking. Retrieved 7/3, 2013, from http://green.dc.gov/energybenchmarking

Decision Tree. (2013). Retrieved 7/2, 2013, from http://en.wikipedia.org/wiki/Decision_tree

Decision Tree learning. (2013). Retrieved 6/20, 2013, from http://en.wikipedia.org/wiki/Decision_tree_learning

Diversity Index. (2013). Retrieved 6/20, 2013, from http://en.wikipedia.org/wiki/Diversity_index

EEBHUB. (2013). Benchmarking and Disclosure. Retrieved 6/4, 2013, from http://www.eebhub.org/research-digest/research-digest-reports/benchmarking-and-disclosure

EnergyIQ. (n.d.). About EnergyIQ. Retrieved 7/1, 2013, from http://energyiq.lbl.gov/EnergyIQ/SupportPages/EIQ-about.jsp

Federspiel, C., Zhang, Q., & Arens, E. (2002). Model-based benchmarking with application to laboratory buildings. *Energy and Buildings, 34*(3), 203-214.

Gini's Diversity Index. (n.d.). Retrieved 6/28, 2013, from http://www.mathworks.com/help/stats/classificationtree.fit.html?searchHighlight=classificationtree

Glazer, J. (2006). ASHRAE 1286-TRP: *Evaluation of Building Energy Performance Rating Protocols.* Park Ridge, Illinois: Project Monitoring Subcommittee - ASHRAE Technical Committee TC 7.6 - Systems Energy Utilization.

Green building alliance - why green building? (2013). Retrieved 7/08, 2013, from http://www.gbapgh.org/content.aspx?ContentID=253

Ho, T. K. (1995). Random decision forests. *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on* (Vol. 1, pp. 278-282). IEEE.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 20*(8), 832-844.

Huang, Y. J., Hanford, J. W., & Piraino, M. (1993). The impact of variations in building parameters and operating conditions on commercial building energy use and load shapes. *Proc. IBPSA Conference,* 16-18.

Identifying outliers. (2013). Retrieved 7/5, 2013, from http://en.wikipedia.org/wiki/Outlier

Inter-quartile range. (2013). Retrieved 7/5, 2013, from http://en.wikipedia.org/wiki/ Interquartile_range

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Ijcai, , 14*(2) 1137-1145.

Linear Regression. (2013). Retrieved 7/2, 2013, from http://en.wikipedia.org/wiki/Linear_regression

Mathew, P., Mills, E., Bourassa, N., & Brook, M. (2008). Action-oriented benchmarking: Using the CEUS database to benchmark commercial buildings in California. *Energy Engineering, 105*(5), 6-18.

Mark Young Training Systems. (n.d.). Retrieved 7/30, 2013, from http://markyoungtrainingsystems.com/2009/08/

Mean squared error. (2013). Retrieved 7/2, 2013, from http://en.wikipedia.org/wiki/Mean_squared_error

Mills, E., Mathew, P., & Piette, M. A. (2008). Action-oriented Benchmarking: Concepts and Tools. *Energy Engineering*, *105*(4), 21-40.

Mistakes to avoid and reporting OLS. (n.d.). Retrieved 7/1, 2013, from http://www.chsbs.cmich.edu/fattah/courses/empirical/29.html

Neida, B. V., & Hicks, T. (2001). Building Performance Defined: the ENERGY STAR National Energy Performance Rating System. Retrieved 7/3, 2013, from http://www.energystar.gov/ia/business/tools_resources/aesp.pdf

Nyce, C., & CPCU, A. (2007). Predictive Analytics White Paper. *American Institute for CPCU/Insurance Institute of America,* 24.

Pacific Northwest National Laboratory. (2008). *Facility Energy Decision System User's Guide: Release 6.0.* Richland: Battelle Memorial Institute.

Pacific Northwest National Laboratory. (2011). Facility Energy Decision System: Home. Retrieved 7/15, 2013, from http://www.pnl.gov/feds/

Partition of Sum of Squares. (2013). Retrieved 7/1, 2013, from http://en.wikipedia.org/wiki/Partition_of_sums_of_squares

Root-mean-square deviation. (2013). Retrieved 7/3, 2013, from http://en.wikipedia.org/wiki/Root_mean_squared_error

Sabapathy, A., Ragavan, S. K. V., Vijendra, M., & Nataraja, A. G. (2010). Energy efficiency benchmarks and the performance of LEED rated buildings for information technology facilities in Bangalore, India. *Energy and Buildings, 42*(11), 2206-2212.

Sharp, T. (1996). Energy Benchmarking in Commercial Office Buildings. *ACEEE 1996 Summer Study on Energy Efficiency in Buildings (4)* (pp. 321-329).

Shmueli, G., Patel, N. R., & Bruce, P. C. (2011). *Data mining for business intelligence: Concepts, techniques, and applications in microsoft office excel with XLMiner.* Wiley. com.

Speybroeck, N. (2012). Classification and regression trees. *International Journal of Public Health, 57*(1), 243-246.

US Department of Energy (DOE). (2008). *Buildings Energy Data Book*. Silver Spring, MD: D&R International, Ltd.

US Energy Information Administration. (2008). Annual Energy Review 2008. Retrieved 7/7, 2013, from http://www.eia.doe.gov/aer/pdf/aer.pdf

US Energy Information Administration. (2012). Annual Energy Outlook 2012. Retrieved 7/5, 2013, from www.eia.gov/pressroom/presentations/howard_01232012.pdf

US Energy Information Administration. (2013). CBECS Status. Retrieved 6/5, 2013, from http://www.eia.gov/consumption/commercial/

US Environmental Protection Agency. (2007). *ENERGY STAR Performance Ratings: Technical Methodology for Office, Bank/Financial Institution, and Courthouse.* Washington DC: US Environmental Protection Agency.

US Environmental Protection Agency. (2007). *Benchmarking to save energy*. Washington, DC: U.S. Environmental Protection Agency.

US Environmental Protection Agency. (2011). ENERGY STAR Portfolio Manager Methodology for Accounting for Weather. Retrieved 7/8, 2013, from http://www.energystar.gov/ia/business/evaluate_performance/Methodology_Weather_20110224.pdf

US Environmental Protection Agency. (2013). *Guidelines for energy management.* Washington, DC: U.S. Environmental Protection Agency.

US Environmental Protection Agency. (n.d.). Learn how Portfolio Manager helps you save. Retrieved 7/10, 2013, from http://www.energystar.gov/buildings/facility-owners-and-managers/existing-buildings/use-portfolio-manager/learn-how-portfolio-manager

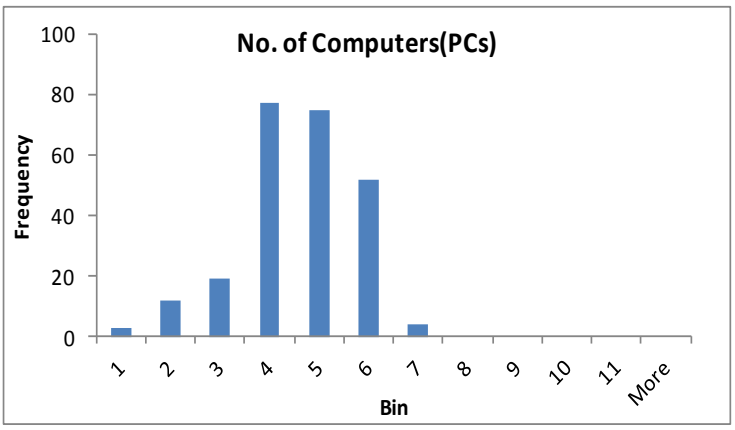
US Environmental Protection Agency. (n.d.). How Portfolio Manager calculates greenhouse gas emissions. Retrieved 7/10, 2013, from http://www.energystar.gov/buildings/facility-owners-and-managers/existing-buildings/use-portfolio-manager/understand-metrics/how

APPENDIX A

DISTRIBUTION OF VARIABLES

The following is the original distribution of variables in the CBECS database for Office buildings:

1. *No. of Personal Computers*



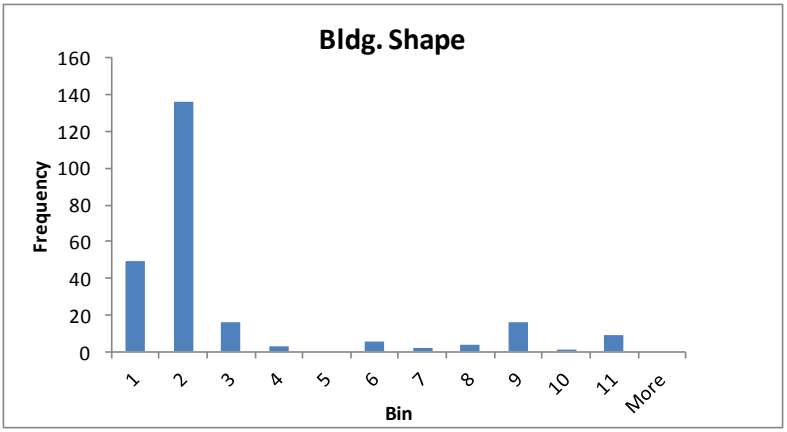
Bin	*Frequency*
1	3
2	12
3	19
4	77
5	75
6	52
7	4
8	0
9	0
10	0
11	0
More	0

Key

1	1 to 4
2	5 to 9
3	10 to 19
4	20 to 49
5	50 to 99
6	100 to 249
7	250 to 499

	Highest frequency of occurrence
	Second highest frequency of occurrence

2. *Building Shape*

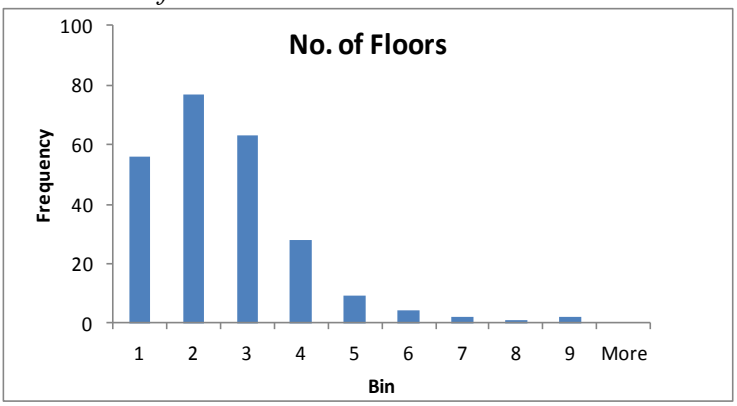


Bin	*Frequency*
1	49
2	136
3	16
4	3
5	0
6	6
7	2
8	4
9	16
10	1
11	9
More	0

Key

1	Square
2	Wide Rectangle
3	Narrow Rectangle
4	Rectangle/Square with courtyard
5	"H" Shaped
6	"U" Shaped
7	"E" Shaped
8	"T" Shaped
9	"L" Shaped
10	"+" Shaped
11	Other

3. No. of Floors

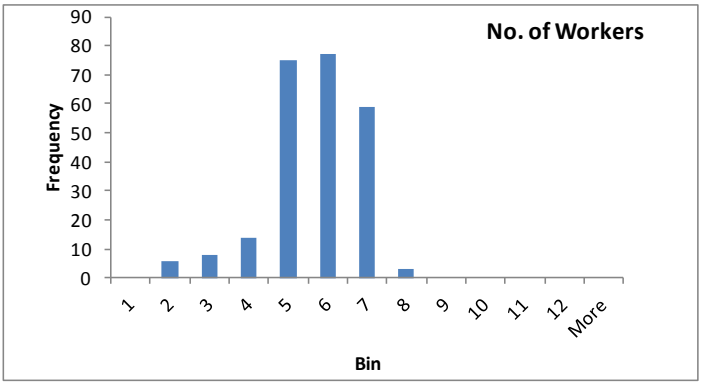


Bin	*Frequency*
1	56
2	77
3	63
4	28
5	9
6	4
7	2
8	1
9	2
More	0

Key

1	1 Story
2	2 Stories
3	3 Stories
4	4 Stories
5	5 Stories
6	6 Stories
7	7 Stories
8	8 Stories
9	9 Stories

4. No. of Workers

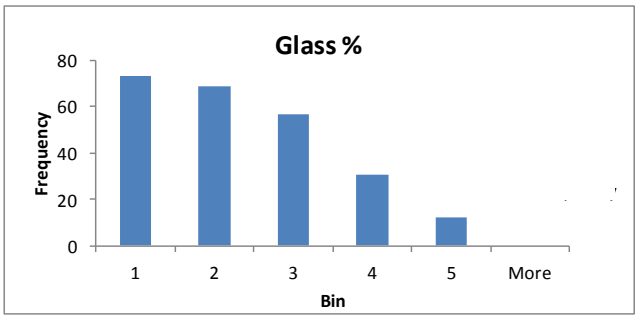


Bin	*Frequency*
1	0
2	6
3	8
4	14
5	75
6	77
7	59
8	3
9	0
10	0
11	0
12	0
More	0

Key

1	None
2	1 to 4
3	5 to 9
4	10 to 19
5	20 to 49
6	50 to 99
7	100 to 249
8	250 to 499
9	500 to 999
10	1000 to 2499
11	2500 to 4999
12	5000 or more

5. Glass percentage

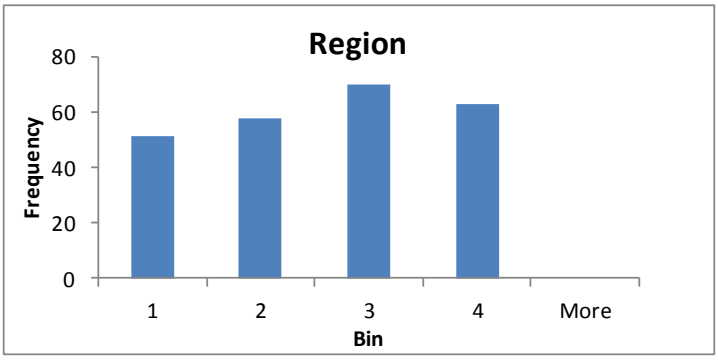


Bin	*Frequency*
1	73
2	69
	57
	31
5	12
More	0

Key

1	10% or less
2	11% to 25%
3	26% to 50%
4	51% to 75%
5	76% to 100%

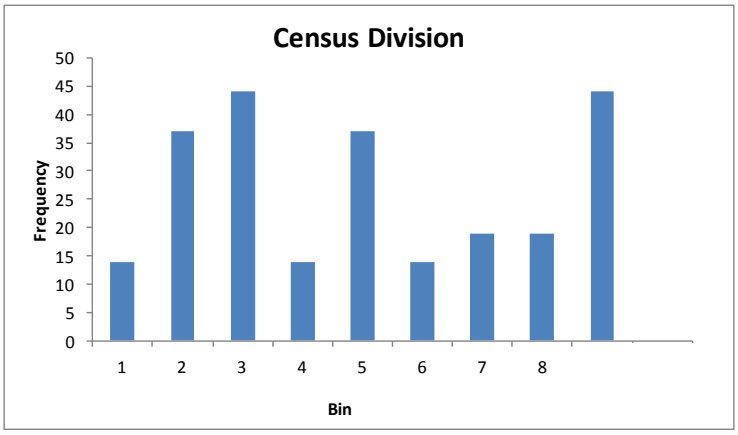
6. *Region*



Bin	*Frequency*
1	51
2	58
3	70
4	63
More	0

KEY	
1	Northeast
2	Midwest
3	South
4	West

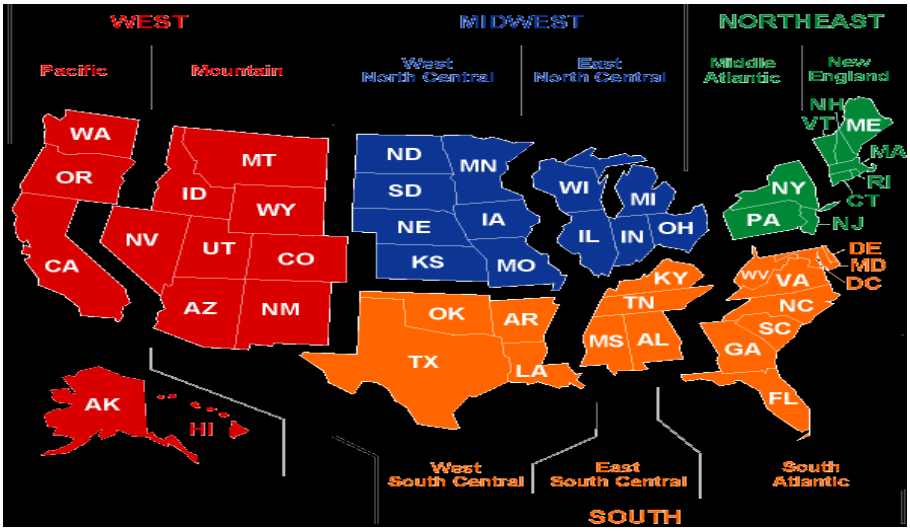
7. *Census Division*



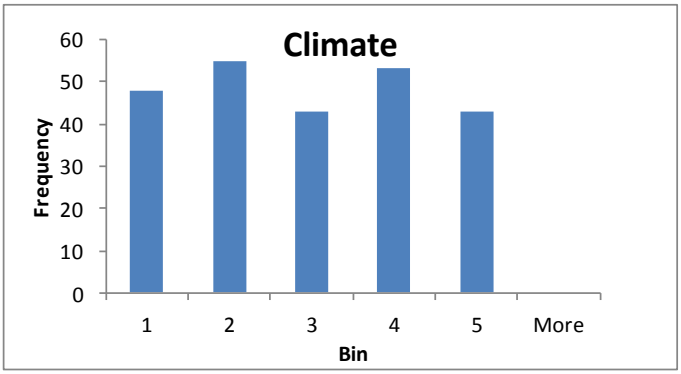
Bin	*Frequency*
1	14
2	37
3	44
4	14
5	37
6	14
7	19
8	19
9	44
More	0

Key

1	New England
2	Middle Atlantic
3	East North Central
4	West North Central
5	South Atlantic
6	East South Central
7	West South Central
8	Mountain
9	Pacific



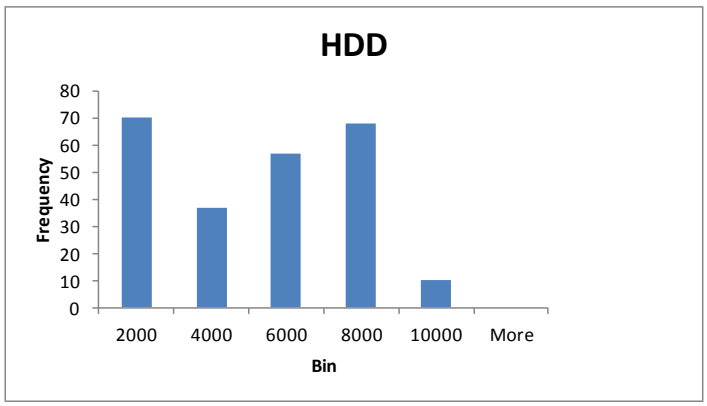
8. Climate



Bin	*Frequency*
1	48
2	55
3	43
4	53
5	43
More	0

KEY	
1	<2000 CDD, >7000 HDD
2	<2000 CDD, 5500-7000 HDD
3	<2000 CDD, 4000-5499 HDD
4	<2000 CDD, <4000 HDD
5	>=2000 CDD, <4000 HDD

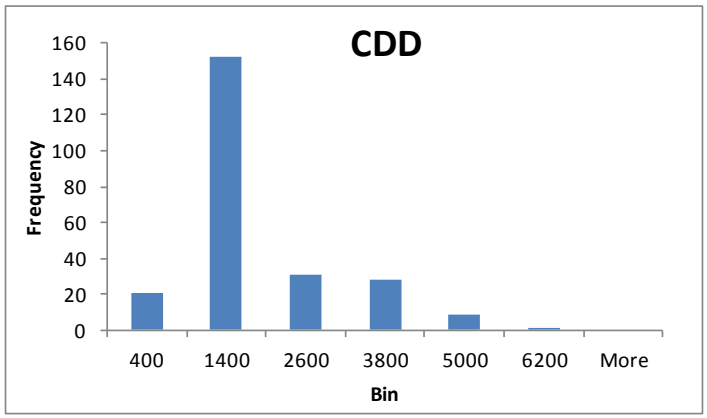
9. HDD



Bin	*Frequency*
2000	70
4000	37
6000	57
8000	68
10000	10
More	0

Note: Heating degree days (HDD) is a numeric value and therefore does not have categories.

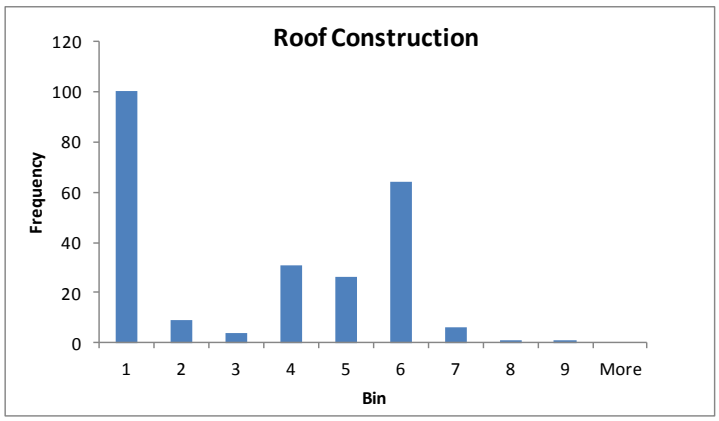
10. CDD



Bin	*Frequency*
400	21
1400	152
2600	31
3800	28
5000	9
6200	1
More	0

Note: Heating degree days (HDD) is a numeric value and therefore does not have categories.

11. Roof Construction

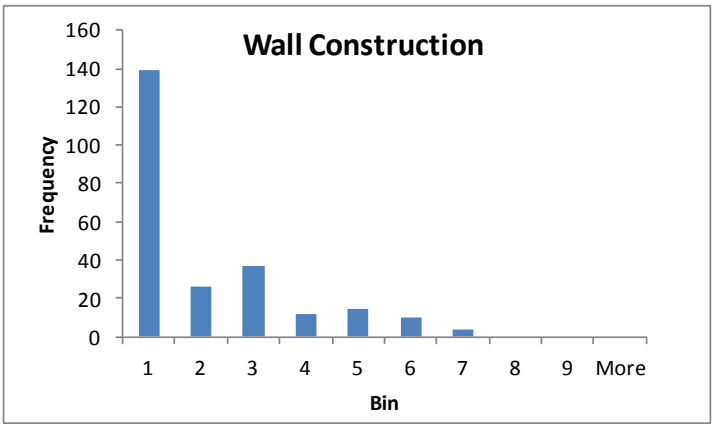


Bin	*Frequency*
1	100
2	9
3	4
4	31
5	26
6	64
7	6
8	1
9	1
More	0

Key

1	Built-Up
2	Slate or Tile shingles
3	Wood Shingles/shakes/othe wood
4	Asphalt/fiberglass/other shingles
5	Metal Surfacing
6	Plastic/rubber/synthetic sheeting
7	Concrete
8	No one major type
9	Other

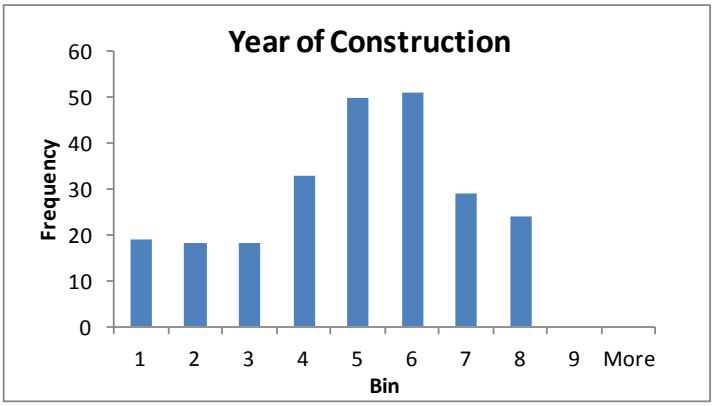
12. Wall Construction



Bin	*Frequency*
1	139
2	26
3	37
4	12
5	14
6	10
7	4
8	0
9	0
More	0

KEY	
1	Brick,Stone or stucco
2	Pre-cast concrete panels
3	Concrete block or poured concrete
4	Siding, Shingles, tiles, or shakes
5	Sheet metal panels
6	Window or vision glass
7	Decorative or construction glass
8	No one major type
9	Other

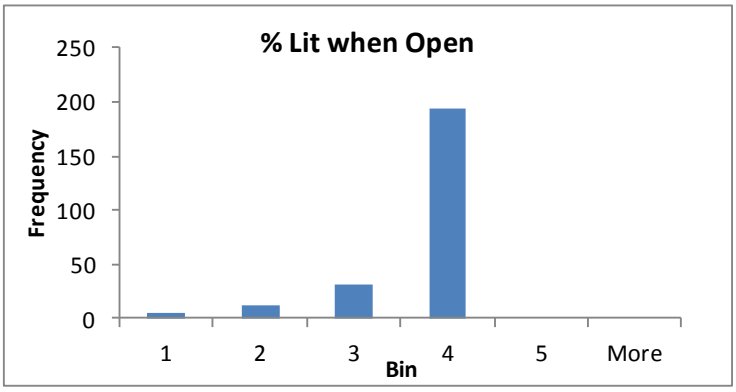
13. Year of Construction



Bin	*Frequency*
1	19
2	18
3	18
4	33
5	50
6	51
7	29
8	24
9	0
More	0

KEY	
1	Before 1920
2	1920 to 1945
3	1946 to 1959
4	1960 to 1969
5	1970 to 1979
6	1980 to 1989
7	1990 to 1999
8	2000 to 2003
9	2004

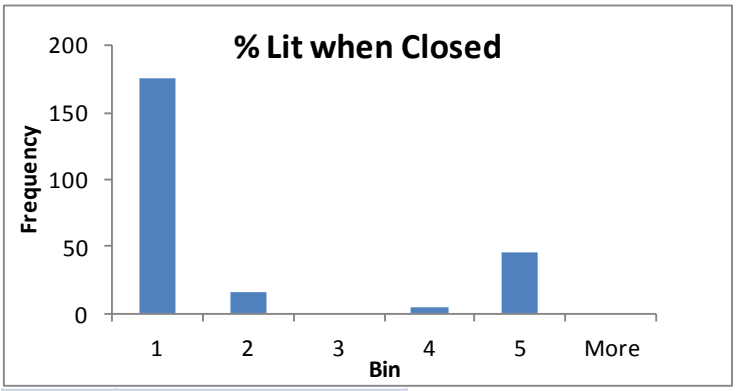
14. Percentage lit when open



Bin	*Frequency*
1	5
2	12
3	32
4	193
5	0
More	0

KEY	
1	1 to 25 %
2	26 to 50%
3	51 to 75%
4	76 to 100 %
5	Not lit at all when open

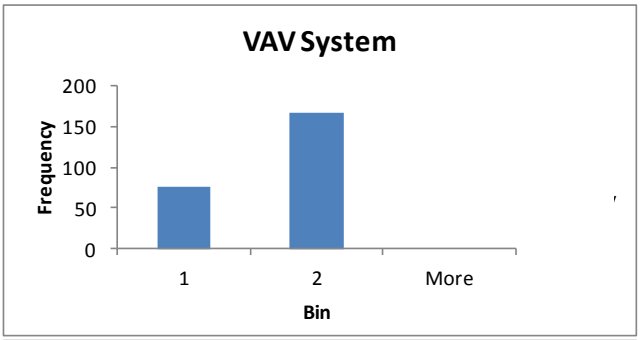
15. Percentage lit when closed



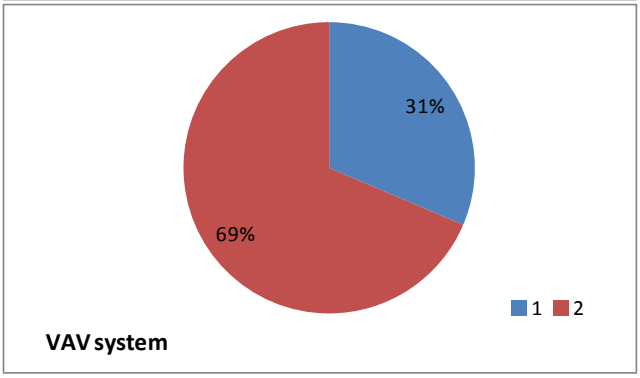
Bin	*Frequency*
1	175
2	16
3	0
4	5
5	46
More	0

KEY	
1	1 to 25 %
2	26 to 50%
3	51 to 75%
4	76 to 100 %
5	Not lit at all when closed

16. VAV (Variable Air Volume)

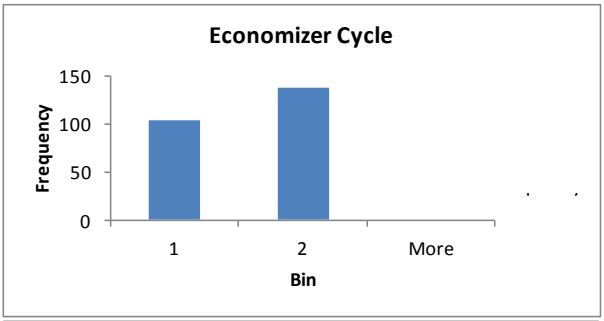


Bin	*Frequency*
1	76
2	166
More	0

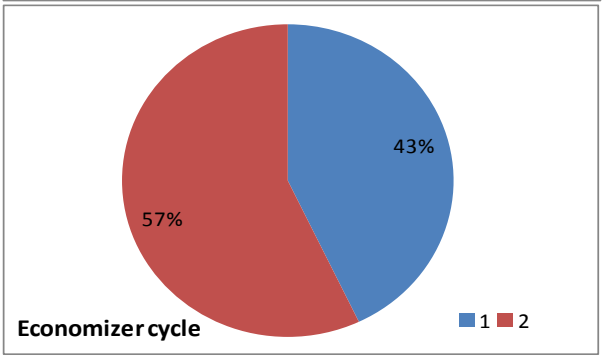


KEY	
1	Yes
2	No

17. Economizer

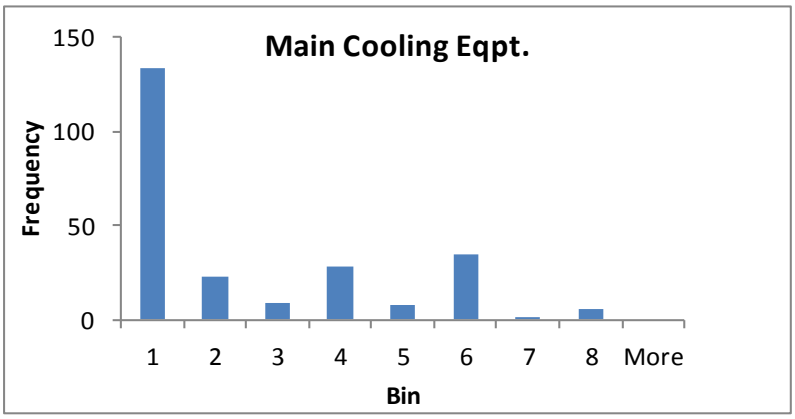


Bin	*Frequency*
1	104
2	138
More	0



KEY	
1	Yes
2	No

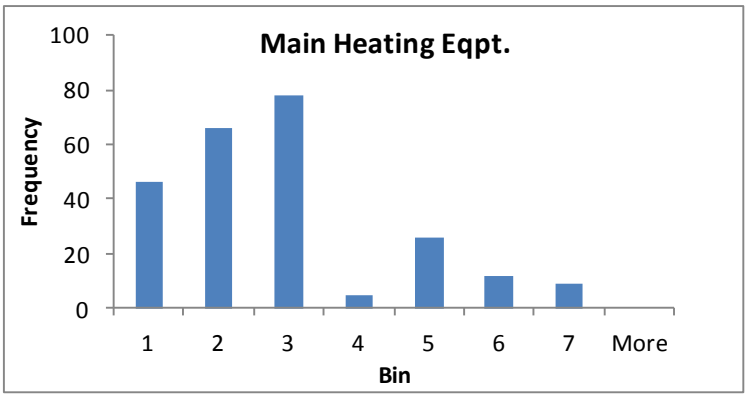
18. Main Cooling Equipment



Bin	*Frequency*
1	133
2	23
3	9
4	28
5	8
6	34
7	1
8	6
More	0

KEY	CLEQP
1	Packaged A/C units
2	Residential-type Central A/C
3	Individual room A/C
4	Heat pumps for cooling
5	District chilled water piped in
6	Central chillers inside the building
7	Evapourative or 'Swamp' coolers
8	Other cooling equipment

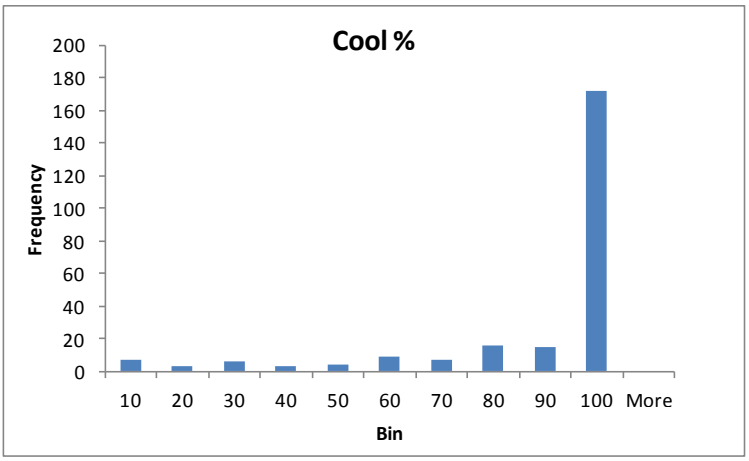
19. Main Heating Equipment



Bin	*Frequency*
1	46
2	66
3	78
4	5
5	26
6	12
7	9
More	0

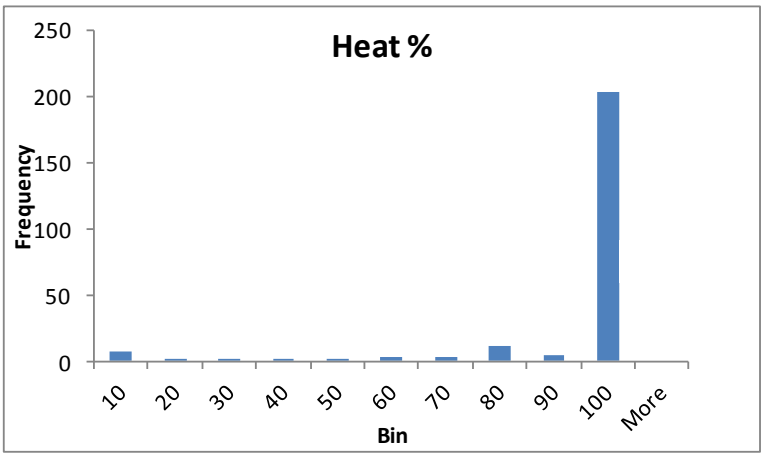
KEY	HTEQP
1	Furnaces that heat air directly
2	Boilers inside the building
3	Packaged heating units
4	Individual space heaters
5	Heat pumps for heating
6	District steam or hot water
7	Other heating equipment

20. Percentage cooled



Bin	*Frequency*
10	7
20	3
30	6
40	3
50	4
60	9
70	7
80	16
90	15
100	172
More	0

21. Percentage heated



Bin	*Frequency*
10	7
20	1
30	1
40	2
50	2
60	4
70	4
80	12
90	5
100	204
More	0

APPENDIX B

DISCRETIZATION

The continuous variables were normalized and then converted to categorical variables.

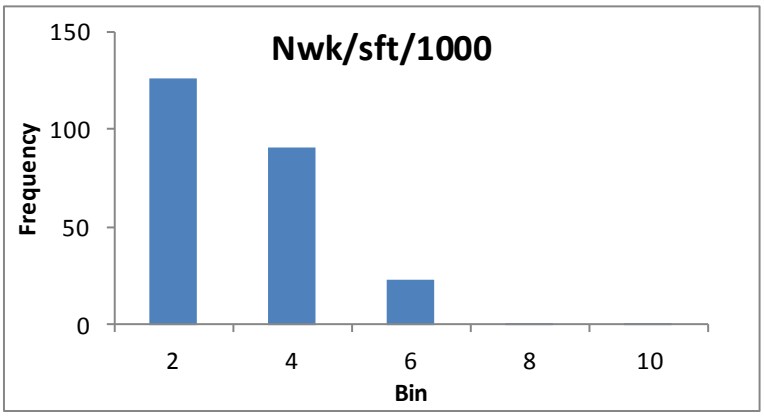
This is given below:

Continuous	Normalization	Categorical
HDDPC	Heating Degree Days*Percent Heated	HDDPCcat
CDDPC	Cooling Degree Days*Percent Cooled	CDDPCcat
NWKR	Number of Workers/(Sqft/1000)	NWKRcat
PCSFT	Number of PCs/(Sqft/1000)	PCSFTcat

The variables were converted to classes based on their distribution.

Office Buildings:

1. *Number of workers/(Sq.ft./1000)*

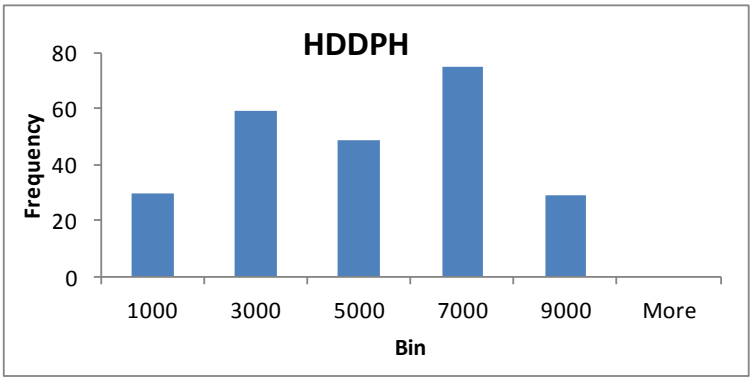


Bin	*Frequency*
2	126
4	91
6	23
8	1
10	1

Nwk/(Sft/1000)	
Minimum	0.02
Maximum	8.22

Range		Class
0	2	1
2.1	4	2
4.1	9	3

2. *Heating degree days*Percent heated*

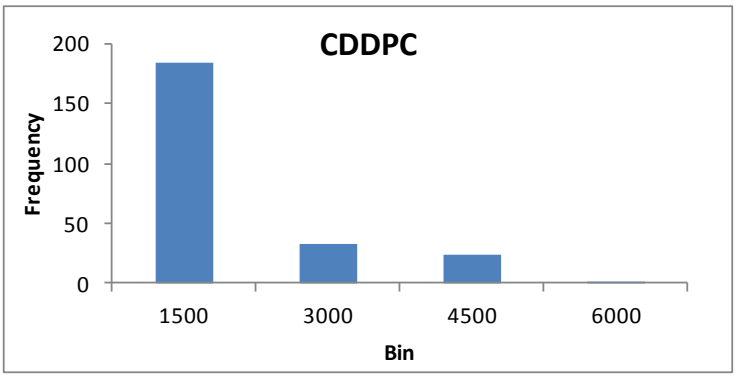


Bin	*Frequency*
1000	30
3000	59
5000	49
7000	75
9000	29

HDDPH	
Minimum	0
Maximum	8968

Range		Class
0	2000	1
2001	4000	2
4001	6000	3
6001	10000	4

3. *Cooling degree days*Percent cooled*

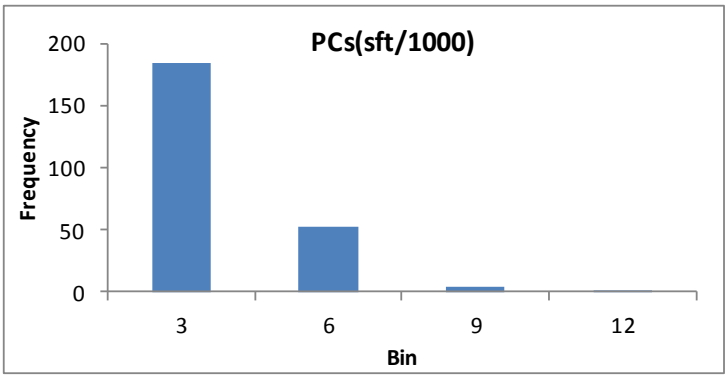


Bin	*Frequency*
1500	184
3000	32
4500	24
6000	2

CDDPC	
Minimum	0
Maximum	5204.7

Range		Class
0	400	1
401	3000	2
3001	6000	3

4. *Number of PCs (Sq.ft/1000)*



Bin	*Frequency*
3	184
6	53
9	4
12	1

PC/(SFT/1000)	
Minimum	0.05
Maximum	11.11

RANGE		CLASS
0	1.5	1
1.6	3	2
3.1	4.5	3
4.6	12	4

School Buildings:

1. *Number of workers/(Sq.ft./1000)*

Nwk(sft/1000)	
Minimum	0.00
Maximum	5.32

Range		Class
0	0.5	1
0.6	2.5	2
2.6	5	3

2. *Heating degree days*Percent heated*

HDD*Percent heated	
Minimum	0
Maximum	9314

Range		Class
0	1400	1
1401	4400	2
4401	7400	3
7401	10400	4

*3. Cooling degree days*Percent cooled*

CDD*Percent cooled	
Minimum	0
Maximum	5654

Range		Class
0	2600	1
2601	4600	2
4601	6600	3

4. Number of PCs(Sq.ft./1000)

PCs(Sqft/1000)	
Minimum	0
Maximum	9.54

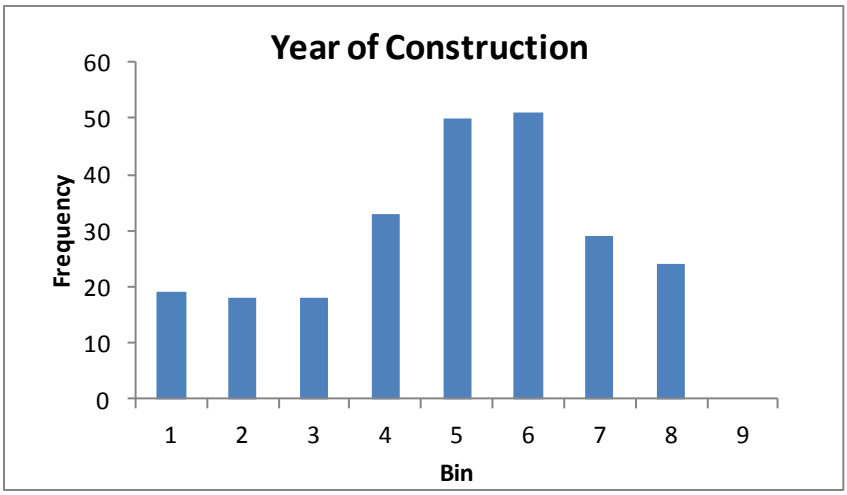
Range		Class
0	3	1
3.1	6	2
6.1	9	3
9.1	12	4

APPENDIX C

REDUCING THE NUMBER OF CATEGORIES

Office Buildings:

1. *Year of construction*

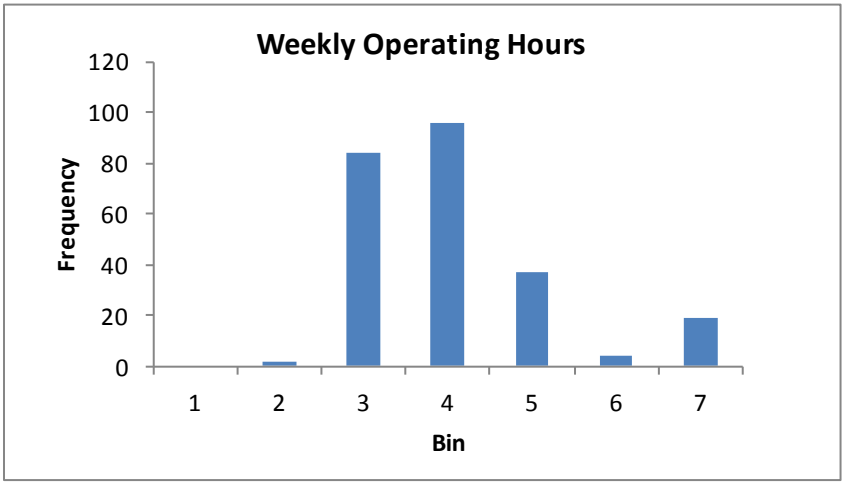


Bin	*Frequency*
1	19
2	18
3	18
4	33
5	50
6	51
7	29
8	24
9	0

KEY	
1	Before 1920
2	1920 to 1945
3	1946 to 1959
4	1960 to 1969
5	1970 to 1979
6	1980 to 1989
7	1990 to 1999
8	2000 to 2003
9	2004

	Collapse into:	
1,2,3	1	Before 1920-till 1959
4,5	2	1960 to 1979
6	3	1980 to 1989
7,8,9	4	1990 to 2004

2. *Weekly operating hours*

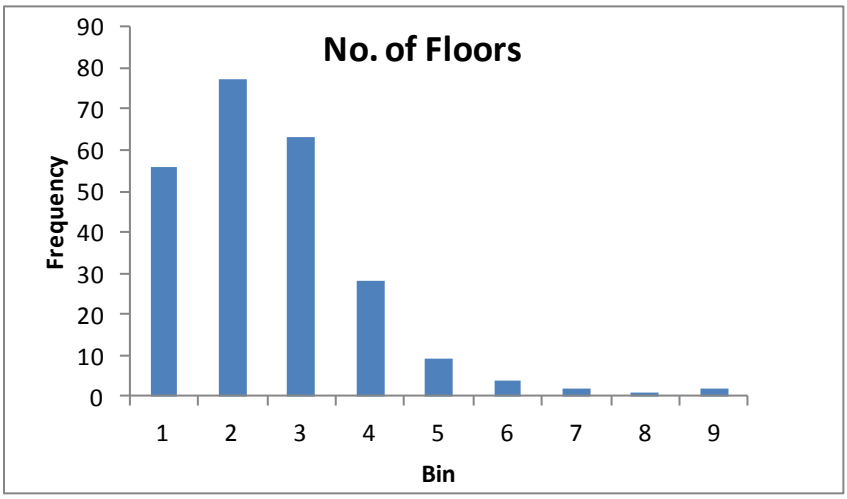


Bin	*Frequency*
1	0
2	2
3	84
4	96
5	37
6	4
7	19

KEY	
1	Zero
2	1 to 39
3	40 to 48
4	49 to 60
5	61 to 84
6	85 to 167
7	Always open

	Collapse to:	
1,2,3	1	1 to 48
4	2	49 to 60
5,6,7	3	61 to 167 and more

3. No. of floors

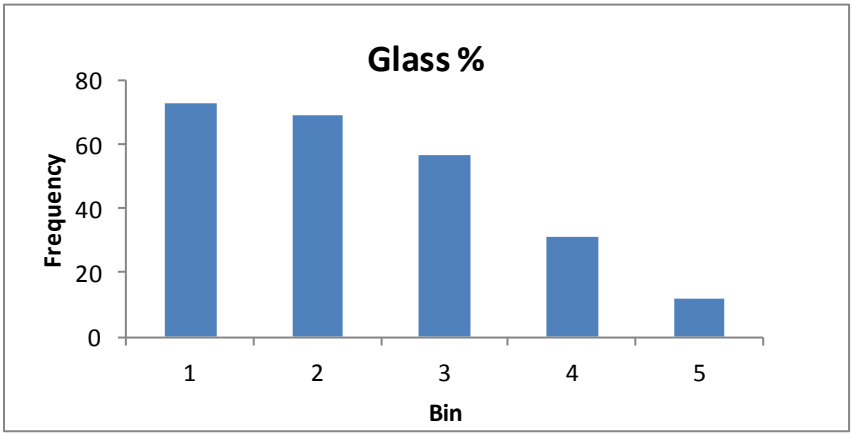


Bin	*Frequency*
1	56
2	77
3	63
4	28
5	9
6	4
7	2
8	1
9	2

KEY	
1	1 Story
2	2 Stories
3	3 Stories
4	4 Stories
5	5 Stories
6	6 Stories
7	7 Stories
8	8 Stories
9	9 Stories

	Collapse to:	
1	1	1 Story
2	2	2 Stories
3	3	3 Stories
4,5,6,7,8,9	4	4 to 9 Stories

4. *Percentage of glass*

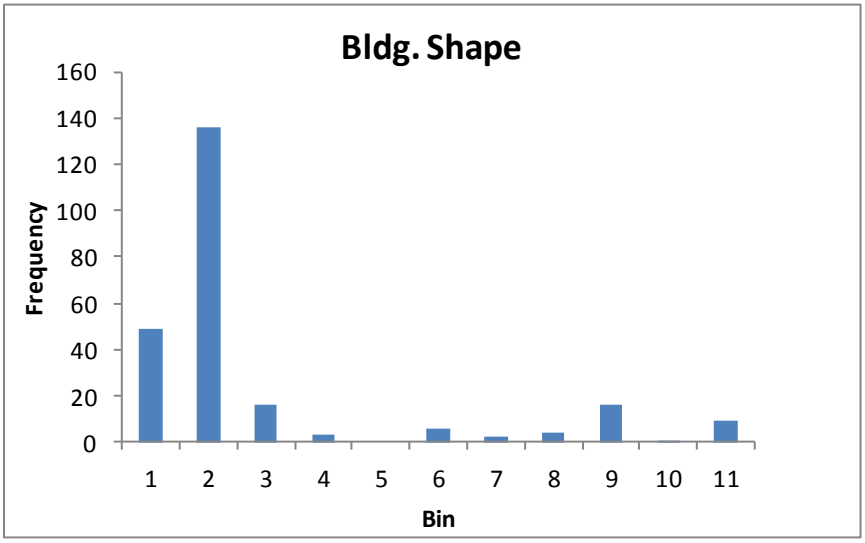


Bin	*Frequency*
1	73
2	69
3	57
4	31
5	12

KEY	
1	10% or less
2	11% to 25%
3	26% to 50%
4	51% to 75%
5	76% to 100%

	Collapse to:	
1	1	10% or Less
2	2	11% to 25%
3	3	26% to 50%
4,5	4	51% to 100%

5. *Building shape*

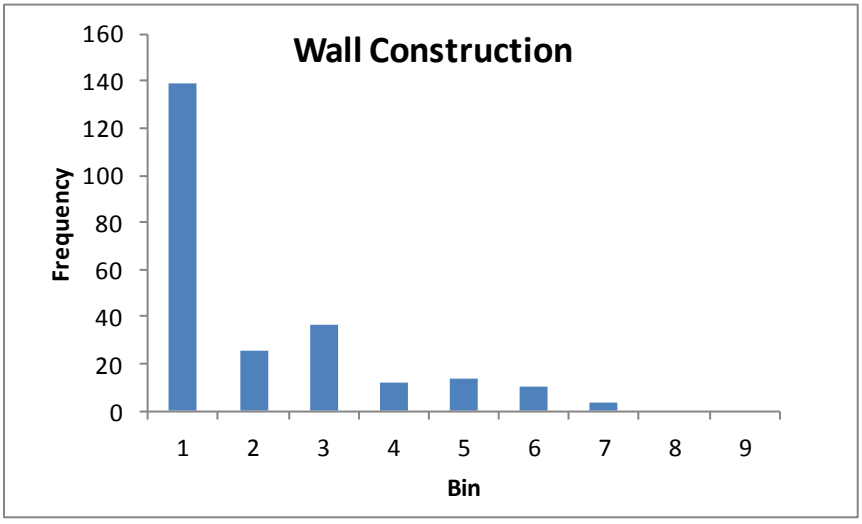


Bin	*Frequency*
1	49
2	136
3	16
4	3
5	0
6	6
7	2
8	4
9	16
10	1
11	9

KEY	
1	Square
2	Wide Rectangle
3	Narrow Rectangle
4	Rectangle/Square with courtyard
5	"H" Shaped
6	"U" Shaped
7	"E" Shaped
8	"T" Shaped
9	"L" Shaped
10	"+" Shaped
11	Other

	Collapse to:	
1	1	Square
2,3	2	Rectangular
4,5,6,7, 8,9	3	Other Shapes

6. Wall construction

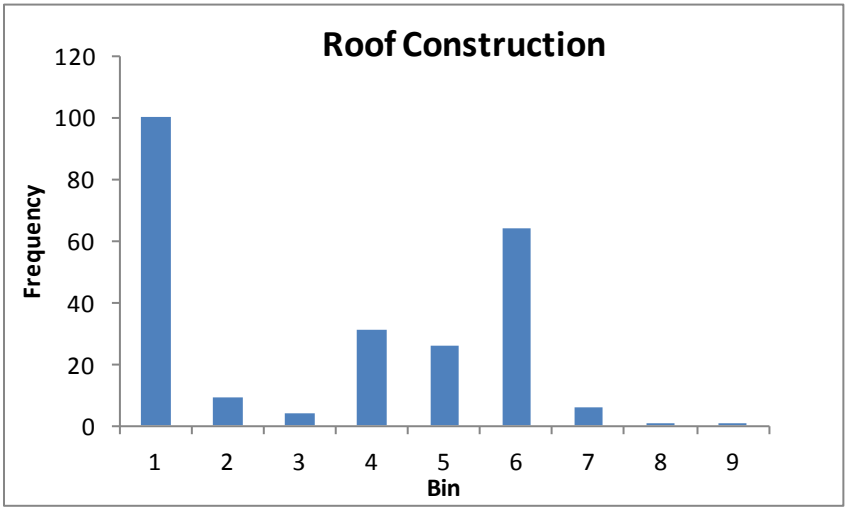


Bin	*Frequency*
1	139
2	26
3	37
4	12
5	14
6	10
7	4
8	0
9	0

KEY	
1	Brick,Stone or stucco
2	Pre-cast concrete panels
3	Concrete block or poured concrete
4	Siding, Shingles, tiles, or shakes
5	Sheet metal panels
6	Window or vision glass
7	Decorative or construction glass
8	No one major type
9	Other

	Collapse to:	
1	1	Brick, Stone or Stucco
2,3	2	Concrete(Pre-cast, Block or poured)
4,5,6,7,8,9	3	Other Types

7. Roof construction

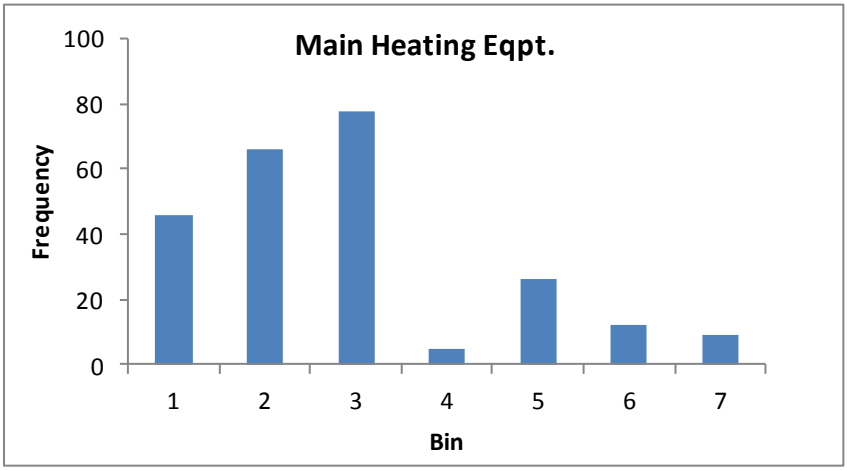


Bin	*Frequency*
1	100
2	9
3	4
4	31
5	26
6	64
7	6
8	1
9	1

KEY	
1	Built-Up
2	Slate or Tile shingles
3	Wood Shingles/shakes/other wood
4	Asphalt/fiberglass/other shingles
5	Metal Surfacing
6	Plastic/rubber/synthetic sheeting
7	Concrete
8	No one major type
9	Other

	Collapse to:	
1	1	Built-Up
2,3,4	2	Slate, Tile, Wood or Other Shingles
5,6	3	Metal/Plastic/Rubber Sheeting
7,8,9	4	Concrete and Other types

8. *Main heating equipment*

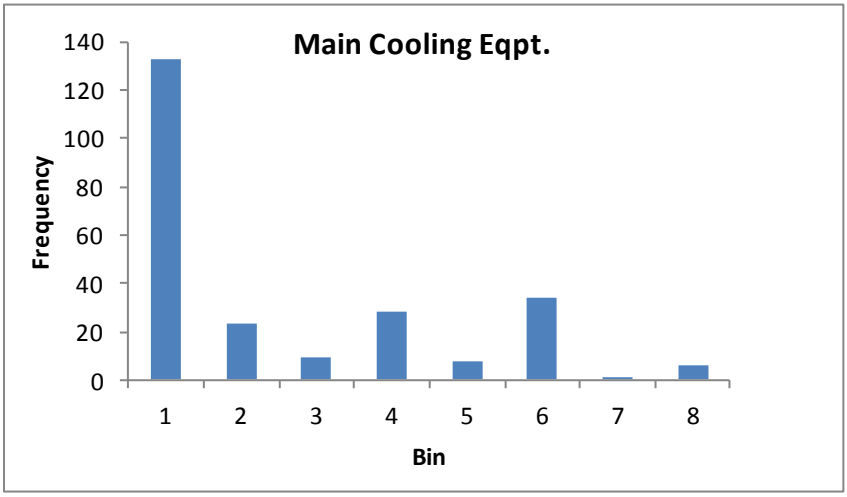


Bin	*Frequency*
1	46
2	66
3	78
4	5
5	26
6	12
7	9

KEY	
1	Furnaces that heat air directly
2	Boilers inside the building
3	Packaged heating units
4	Individual space heaters
5	Heat pumps for heating
6	District steam or hot water
7	Other heating equipment

	Collapse to:	
1	1	Furnaces
2,6	2	Boilers/District Steam or Hot waters
3,4	3	Package Units and Individual Space heaters
5,7	4	Heat Pumps and Other types

9. *Main cooling equipment*

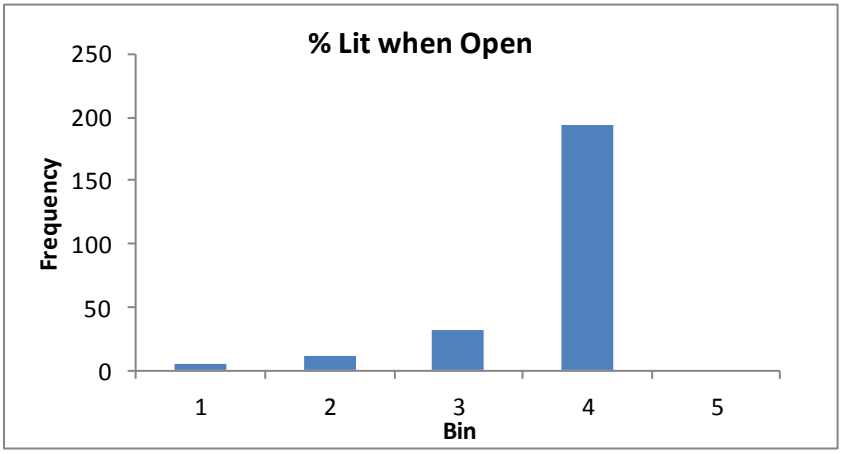


Bin	*Frequency*
1	133
2	23
3	9
4	28
5	8
6	34
7	1
8	6

KEY	
1	Packaged A/C units
2	Residential-type Central A/C
3	Individual room A/C
4	Heat pumps for cooling
5	District chilled water piped in
6	Central chillers inside the building
7	Evapourative or 'Swamp' coolers
8	Other cooling equipment

	Collapse to:	
1,3	1	Packaged A/C and Individual room
2,4	2	Central-Residential or central chillers
5,6	3	Chilled Water and Heat Pump
7,8	4	Evapourative and Other types

10. Percentage lit when open

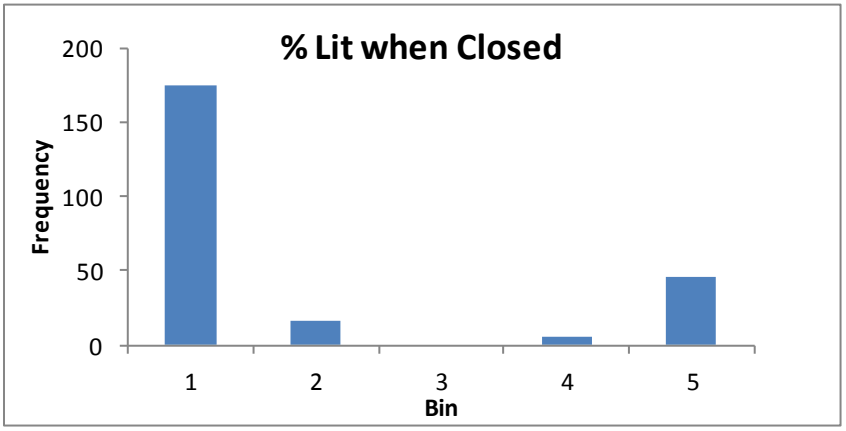


Bin	*Frequency*
1	5
2	12
3	32
4	193
5	0

KEY	
1	1 to 25 %
2	26 to 50%
3	51 to 75%
4	76 to 100 %
5	Not lit at all when open

	Collapse to:	
1,2	1	1 to 50%
3	2	51 to 75%
4,5	3	76% to 100%

11. Percentage lit when closed



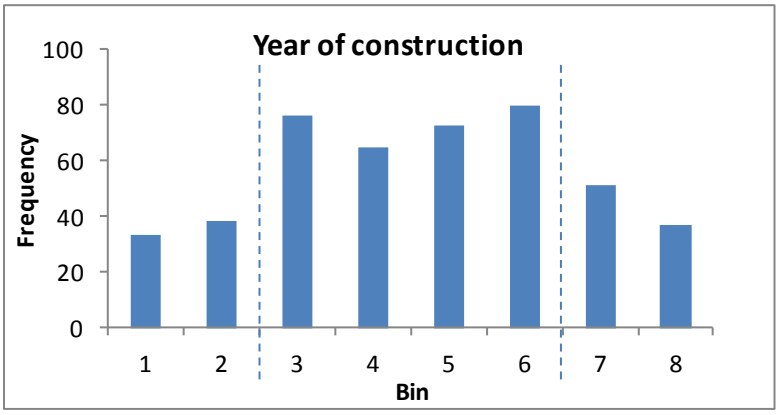
Bin	*Frequency*
1	175
2	16
3	0
4	5
5	46

KEY	
1	1 to 25 %
2	26 to 50%
3	51 to 75%
4	76 to 100 %
5	Not lit at all when closed

	Collapse to:	
1	1	1 to 25%
2,3,4	2	26 to 100%
5	3	Not lit at all when closed

Combined dataset:

1. *Year of construction*

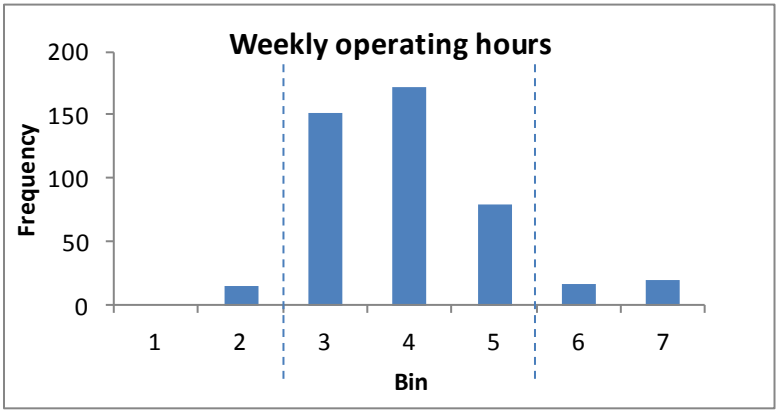


Bin	*Frequency*
1	33
2	38
3	76
4	65
5	73
6	80
7	51
8	37

KEY	
1	Before 1920
2	1920 to 1945
3	1946 to 1959
4	1960 to 1969
5	1970 to 1979
6	1980 to 1989
7	1990 to 1999
8	2000 to 2003
9	2004

	Collapse to:	
1,2	1	Before 1920-till 1959
3,4,5,6	2	1960 to 1989
7,8,9	3	1990 to 2004

2. *Weekly operating hours*

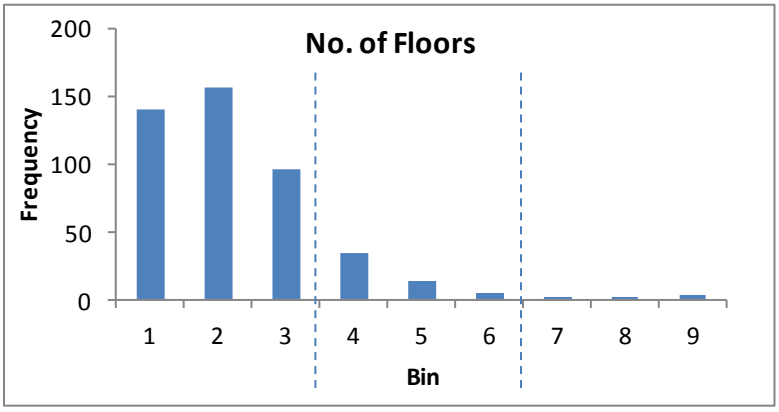


Bin	*Frequency*
1	0
2	15
3	151
4	172
5	79
6	16
7	20

KEY	
1	Zero
2	1 to 39
3	40 to 48
4	49 to 60
5	61 to 84
6	85 to 167
7	Always open

	Collapse to:	
1,2	1	1 to 39
3,4,5	2	40 to 84
6,7	3	85 to 167 and more

3. No. of floors

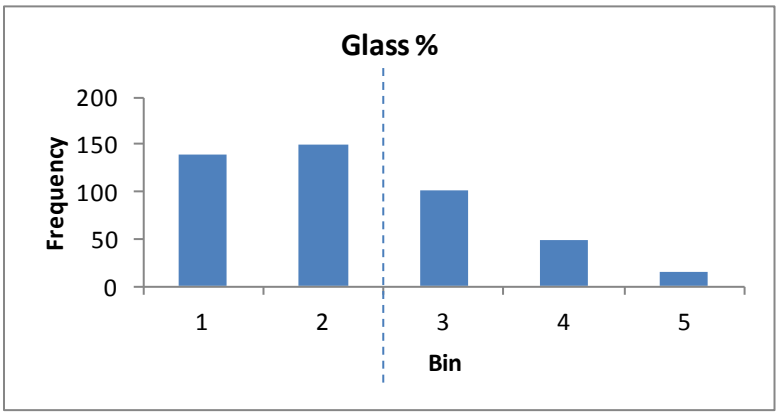


Bin	*Frequency*
1	141
2	156
3	97
4	34
5	14
6	5
7	2
8	1
9	3

KEY	
1	1 Story
2	2 Stories
3	3 Stories
4	4 Stories
5	5 Stories
6	6 Stories
7	7 Stories
8	8 Stories
9	9 Stories

	Collapse to:	
1,2,3	1	1,2,3 Stories
4,5,6	2	4,5,6 Stories
7,8,9	3	7,8,9 Stories

4. Percentage of glass

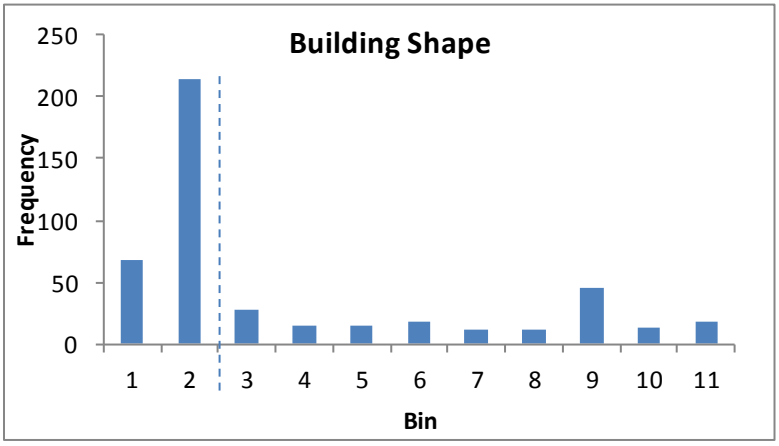


Bin	*Frequency*
1	139
2	149
3	102
4	49
5	14

KEY	
1	10% or less
2	11% to 25%
3	26% to 50%
4	51% to 75%
5	76% to 100%

	Collapse to:	
1,2	1	10% to 25%
3,4,5	2	26% to 100%

5. *Building shape*

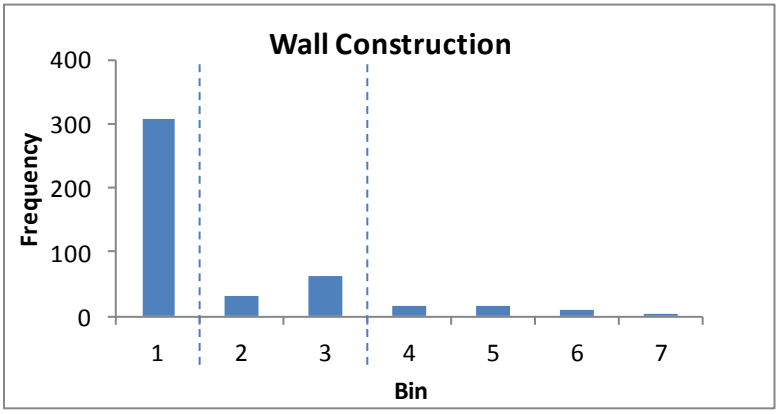


Bin	*Frequency*
1	67
2	214
3	27
4	14
5	14
6	18
7	11
8	12
9	45
10	13
11	18

KEY	
1	Square
2	Wide Rectangle
3	Narrow Rectangle
4	Rectangle/Square with courtyard
5	"H" Shaped
6	"U" Shaped
7	"E" Shaped
8	"T" Shaped
9	"L" Shaped
10	"+" Shaped
11	Other

	Collapse to:	
1	1	Square
2,3	2	Rectangular
4,5,6,7, 8,9	3	Other Shapes

6. Wall construction

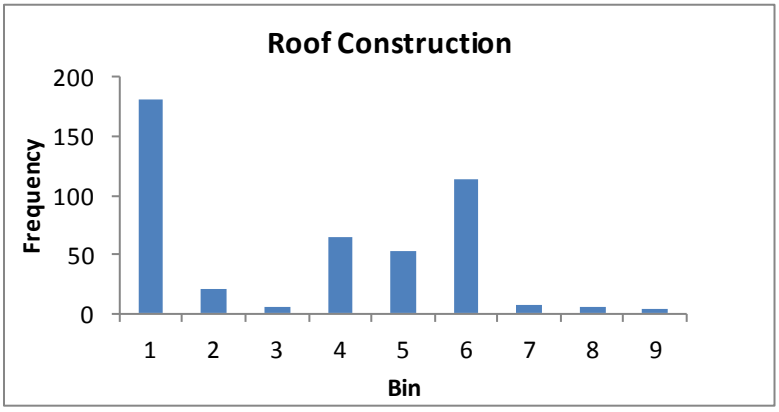


Bin	*Frequency*
1	307
2	32
3	64
4	17
5	17
6	11
7	5

KEY	
1	Brick,Stone or stucco
2	Pre-cast concrete panels
3	Concrete block or poured concrete
4	Siding, Shingles, tiles, or shakes
5	Sheet metal panels
6	Window or vision glass
7	Decorative or construction glass
8	No one major type
9	Other

	Collapse to:	
1	1	Brick, Stone or Stucco
2,3	2	Concrete(Pre-cast, Block or poured)
4,5,6,7,8,9	3	Other Types

7. Roof construction

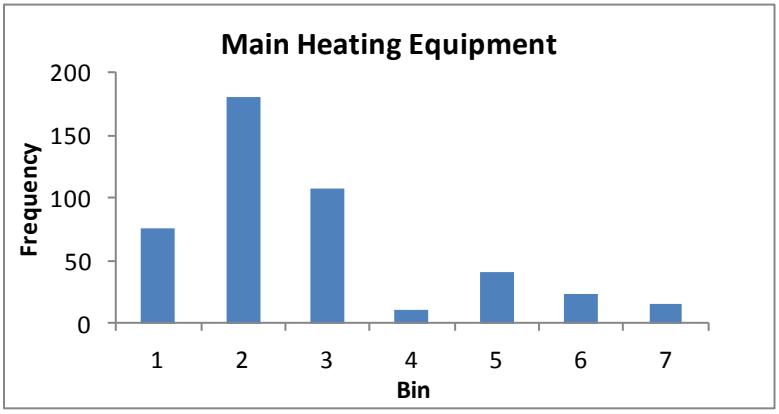


Bin	*Frequency*
1	181
2	21
3	5
4	65
5	52
6	113
7	8
8	5
9	3

KEY	
1	Built-Up
2	Slate or Tile shingles
3	Wood Shingles/shakes/other wood
4	Asphalt/fiberglass/other shingles
5	Metal Surfacing
6	Plastic/rubber/synthetic sheeting
7	Concrete
8	No one major type
9	Other

	Collapse to:	
1	1	Built-Up
2,3,4	2	Slate, Tile, Wood or Other Shingles
5,6	3	Metal/Plastic/Rubber Sheeting
7,8,9	4	Concrete and Other types

8. *Main heating equipment*

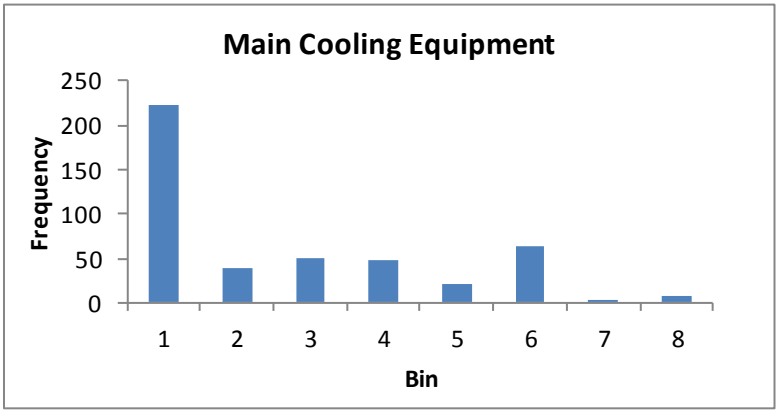


Bin	*Frequency*
1	75
2	180
3	108
4	11
5	41
6	23
7	15

KEY	
1	Furnaces that heat air directly
2	Boilers inside the building
3	Packaged heating units
4	Individual space heaters
5	Heat pumps for heating
6	District steam or hot water
7	Other heating equipment

	Collapse to:	
1	1	Furnaces
2,6	2	Boilers/District Steam or Hot waters
3,4	3	Package Units and Individual Space heaters
5,7	4	Heat Pumps and Other types

9. Main cooling equipment

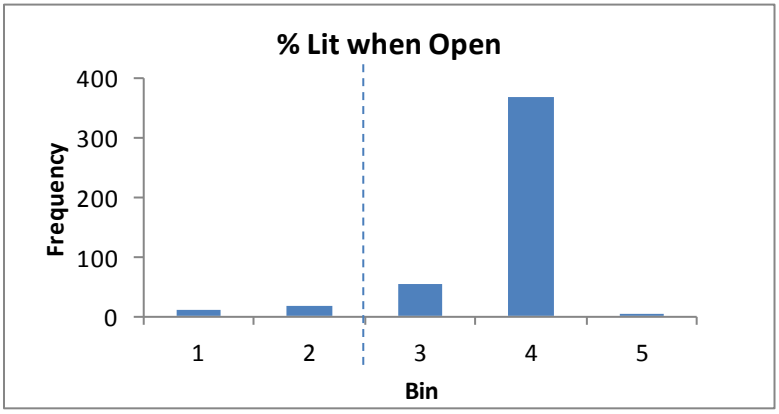


Bin	*Frequency*
1	223
2	38
3	49
4	48
5	21
6	64
7	3
8	7

KEY	
1	Packaged A/C units
2	Residential-type Central A/C
3	Individual room A/C
4	Heat pumps for cooling
5	District chilled water piped in
6	Central chillers inside the building
7	Evapourative or 'Swamp' coolers
8	Other cooling equipment

	Collapse to:	
1,3	1	Packaged A/C and Individual room
2,4	2	Central-Residential or central chillers
5,6	3	Chilled Water and Heat Pump
7,8	4	Evapourative and Other types

10. Percentage lit when open

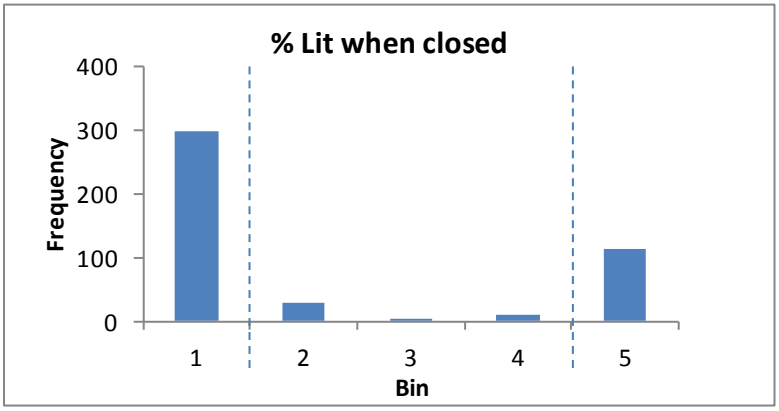


Bin	*Frequency*
1	11
2	19
3	55
4	367
5	1

KEY	
1	1 to 25 %
2	26 to 50%
3	51 to 75%
4	76 to 100 %
5	Not lit at all when open

	Collapse to:	
1,2	1	1 to 50%
3	2	51 to 75%
4,5	3	76% to 100%

11. Percentage lit when closed

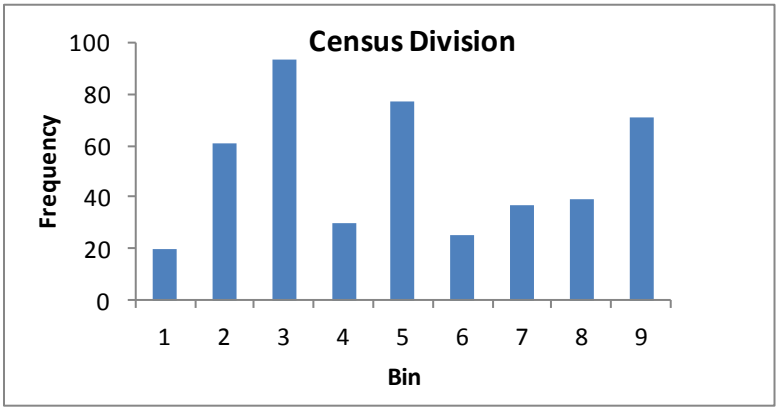


Bin	*Frequency*
1	297
2	28
3	3
4	10
5	115

KEY	
1	1 to 25 %
2	26 to 50%
3	51 to 75%
4	76 to 100 %
5	Not lit at all when closed

	Collapse to:	
1	1	1 to 25%
2,3,4	2	26 to 100%
5	3	Not lit at all when closed

12. Census division



Bin	*Frequency*
1	20
2	61
3	93
4	30
5	77
6	25
7	37
8	39
9	71

Note: The categories for this variable were retained as in the original database since each represents a location in the US specified by CBECS